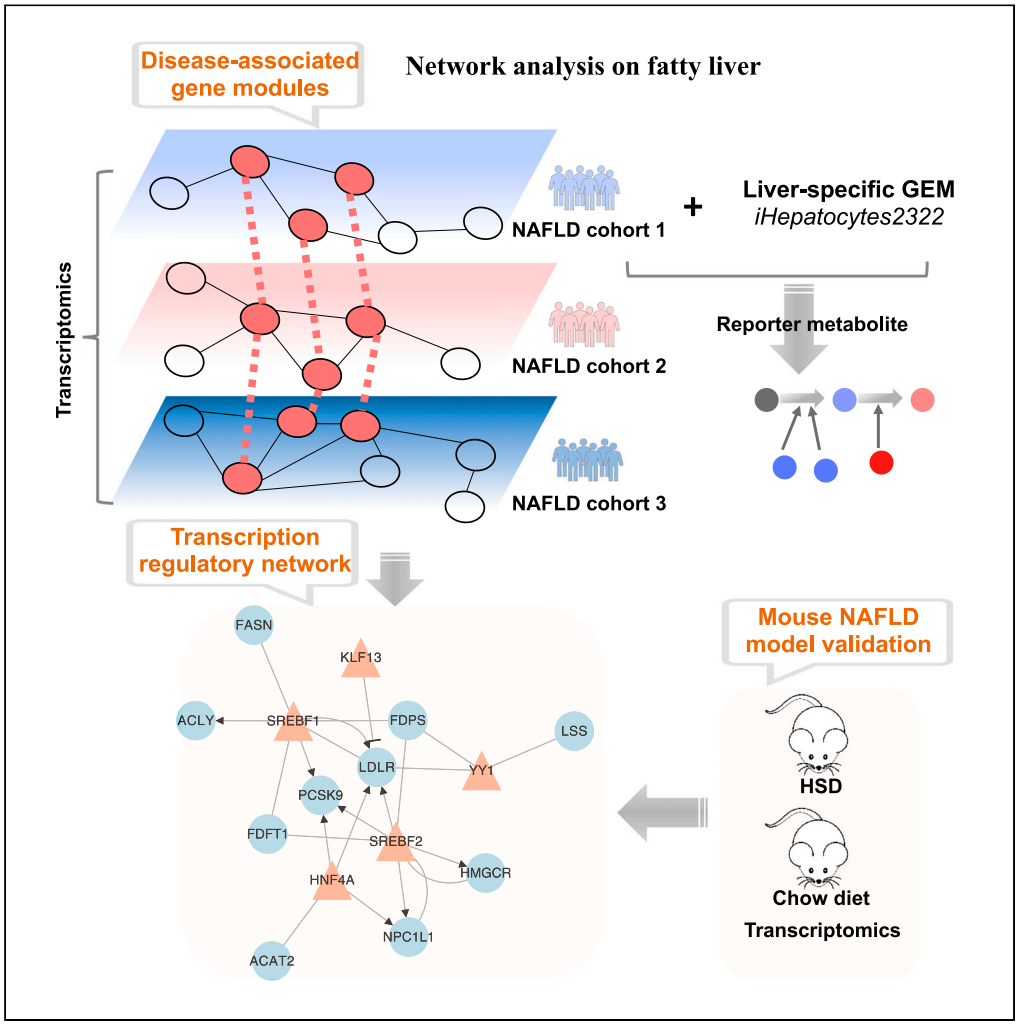
A network-based approach reveals the dysregulated transcriptional regulation in non-alcoholic liver disease

iScience

CellPress
OPEN ACCESS

Article

A network-based approach reveals the dysregulated transcriptional regulation in non-alcoholic fatty liver disease



Hong Yang, Muhammad Arif, Meng Yuan, ..., Jan Borén, Cheng Zhang, Adil Mardinoglu

cheng.zhang@scilifelab.se (C.Z.)
adilm@scilifelab.se (A.M.)

Highlights

Disease-associated gene modules are conserved across multiple NAFLD cohorts

The central genes in disease-associated modules are key enzymes in cholesterol synthesis

YY1 and KLF13 are potential key transcriptional regulators of NAFLD development

Yang et al., iScience 24, 103222
November 19, 2021
https://doi.org/10.1016/j.isci.2021.103222

iScience

CellPress
OPEN ACCESS

Article

A network-based approach reveals the dysregulated transcriptional regulation in non-alcoholic fatty liver disease

Hong Yang,[1] Muhammad Arif,[1] Meng Yuan,[1] Xiangyu Li,[1] Koeun Shong,[1] Hasan Türkez,[2] Jens Nielsen,[3,4] Mathias Uhlén,[1] Jan Borén,[5] Cheng Zhang,[1,6,*] and Adil Mardinoglu[1,7,8,*]

SUMMARY

Non-alcoholic fatty liver disease (NAFLD) is a leading cause of chronic liver disease worldwide. We performed network analysis to investigate the dysregulated biological processes in the disease progression and revealed the molecular mechanism underlying NAFLD. Based on network analysis, we identified a highly conserved disease-associated gene module across three different NAFLD cohorts and highlighted the predominant role of key transcriptional regulators associated with lipid and cholesterol metabolism. In addition, we revealed the detailed metabolic differences between heterogeneous NAFLD patients through integrative systems analysis of transcriptomic data and liver-specific genome-scale metabolic model. Furthermore, we identified transcription factors (TFs), including SREBF2, HNF4A, SREBF1, YY1, and KLF13, showing regulation of hepatic expression of genes in the NAFLD-associated modules and validated the TFs using data generated from a mouse NAFLD model. In conclusion, our integrative analysis facilitates the understanding of the regulatory mechanism of these perturbed TFs and their associated biological processes.

INTRODUCTION

Non-alcoholic fatty liver disease (NAFLD) is considered as one of the most important causes of liver disease, worldwide (Asrani et al., 2019). The global prevalence of NAFLD was estimated to be 25% and has increased rapidly (Huang et al., 2021; Younossi et al., 2016, 2018). NAFLD is characterized by the hepatic accumulation of triglycerides, spanning from simple non-alcoholic fatty liver (NAFL) to non-alcoholic steatohepatitis (NASH) that might progress to cirrhosis and hepatocellular carcinoma (HCC) (Friedman et al., 2018; Huang et al., 2021; Ioannou et al., 2019). Moreover, NAFLD is strongly associated with obesity, diabetes, and cardiovascular disease, therefore, it has drastically increased in patient groups with these diseases (Golabi et al., 2019; Ye et al., 2020; Younossi et al., 2019). Despite the high degree of popularity, no effective therapies are yet approved for the treatment of NAFLD (El-Agroudy et al., 2019; Mullard, 2020; Newsome et al., 2021; Stower, 2021). Hence, a comprehensive understanding of the underlying molecular mechanism of NAFLD is critical for the development of novel approaches for its prevention and treatment.

Biological networks provide a robust framework for integrating omics data, elucidating pathophysiological responses, and revealing the underlying molecular mechanisms involved in the progression of disease (Calabrese et al., 2017; Mardinoglu et al., 2018; Nayak et al., 2009). Biological networks, including protein-protein interaction networks, transcriptional regulatory networks (RNs), gene co-expression networks (GCNs), genome-scale metabolic models (GEMs) and integrated networks (INs), are widely used in systems analysis (Mardinoglu et al., 2018). The central goal of biological network analysis is to identify critical functional units (so-called modules) and their constituent genes (Califano et al., 2012; Choobdar et al., 2019). By investigating the importance of these functional modules in disease pathogenesis, it is possible to understand the biological mechanisms that underpin the disease (Cerami et al., 2010; Huan et al., 2013; Wainberg et al., 2021). In particular, GCNs of 17 human cancers and 46 human tissues have been generated and used to gain insights into disease mechanisms by identifying the key biological components of the cancers or tissues (Arif et al., 2021; Lee et al., 2018; Uhlen et al., 2017). GEMs are reconstructed by incorporating all biochemical reactions and transport processes in a cell or tissue and have been extensively used to

[1]Science for Life Laboratory, KTH - Royal Institute of Technology, Stockholm, Sweden

[2]Department of Medical Biology, Faculty of Medicine, Atatürk University, Erzurum, Turkey

[3]Department of Biology and Biological Engineering, Chalmers University of Technology, 41296 Gothenburg, Sweden

[4]BioInnovation Institute, 2200 Copenhagen, Denmark

[5]Department of Molecular and Clinical Medicine, University of Gothenburg and Sahlgrenska University Hospital, Gothenburg, Sweden

[6]School of Pharmaceutical Sciences, Zhengzhou University, Zhengzhou, PR China

[7]Centre for Host-Microbiome Interactions, Faculty of Dentistry, Oral & Craniofacial Sciences, King's College London, London, UK

[8]Lead contact

*Correspondence: cheng.zhang@scilifelab.se (C.Z.), adilm@scilifelab.se (A.M.)

https://doi.org/10.1016/j.isci.2021.103222

discover potential biomarkers and drug targets, as well as to reveal the mode of action of a drug (Lewis and Kemp, 2021; Mardinoglu et al., 2018).

To date, GCNs have been used for investigating the causal mechanisms underlying NAFLD using mouse population data (Chella Krishnan et al., 2018) and human population data (Zhang et al., 2020) and for integrative analysis of mouse model data and patients data (Saeed, 2021). However, there remains a lack of holistic studies of samples that cover a large spectrum of disease severity. Moreover, the heterogeneity of clinical manifestations among NAFLD patients acts as an essential impediment for the discovery of critical pathogenic drivers, requiring in-depth systematic analysis consequently (Alonso et al., 2017). Recently, several liver-biopsy proved transcriptomics data from large patients' cohorts had been conducted (Azzu et al., 2021; Govaere et al., 2020; Hoang et al., 2019), and these datasets may be used to provide significant functional insights based on network analysis that cannot be derived from individual gene-level analysis.

In this study, we employed an integrative systems biology approach by integrating NAFLD transcriptomics data with biological networks and elucidated the molecular mechanisms underlying NAFLD progression. We first generated GCNs for liver tissue of normal and NAFLD patients based on transcriptomics data and identified the perturbed modules associated with the severity of NAFLD. Secondly, we employed a liver-specific GEM called *iHepatocytes2322* to analyze the differential expression data, and gained insights into detailed metabolic differences in NAFLD. Third, we subsequently validated the perturbed modules using transcriptomics data from another two independent studies and highlighted the disease-associated modules that are conserved across multiple NAFLD cohorts by combing functional and topological similarities. Next, we used a liver cancer data set in The Cancer Genome Atlas (TCGA) to investigate if the dysregulated expression of genes in the disease-associated modules is relevant to patient outcome. Finally, we performed transcription factor (TF)-target regulatory network analysis, identified TFs that regulate disease-associated modules, and validated those TFs with the transcriptomics data from a mouse NAFLD model fed by high-sucrose diet (HSD).

RESULTS

Generation of co-expression networks for liver tissue

To identify the robust co-expressed genes showing transcriptional differences between the liver tissue of normal subjects and NAFLD patients, we first constructed GCN of 226 non-diseased liver tissue based on the transcriptomics data from the Genotype-Tissue Expression (GTEx) database (GTEx analysis V8) (Consortium, 2013) and GCN of liver tissue from NAFLD cohort including 10 normal samples, 50 patients with NAFL, and 155 patients with NASH (Govaere et al., 2020) (Figure 1A). We filtered out lowly expressed genes for each data set based on their mean gene expression level (TPM < 1) and performed Spearman's rank correlation test between each gene pair. All p values were adjusted by FDR correction (Benjamini–Hochberg). Afterward, we retained gene pairs with top 10% significantly positive correlation (FDR < 0.05) on the networks (Figure 1A) and used the Leiden algorithm (Traag et al., 2019) to identify modules of genes from the network. In total, Leiden graph-based clustering identified six and five modules of genes in the GTEx cohort and NAFLD cohort, respectively. Each module in the same cohort consists of uniquely assigned genes with a substantial similarity between gene expression profile (Figure 1B; Data S1). Of note, we found that gene members of any module in the GTEx cohort were different from that of modules in the NAFLD cohort even though 95.9% genes comprising modules in the NAFLD cohort were included by GTEx modules (Figure 1C).

Identification of perturbed modules in NAFLD

We investigated whether the differences in module composition correlated with the molecular changes underlying NAFLD progression. We first identified differentially expressed genes (DEGs) to reveal the global transcriptomic differences in the liver of patients with NAFLD. We observed that 3,167 and 3,285 genes were significantly upregulated (adjusted p value < 0.01) between NAFL and control samples and between NASH and control samples, respectively (Figure 1D; Data S2). Enrichment analysis in KEGG pathway showed that the upregulated DEGs are mostly enriched in the pathways associated with endocytosis, axon guidance, adherens junction, insulin resistance, and insulin signaling (Figure 1E; Data S3). Moreover, we found that 3,176 and 3,140 genes were significantly downregulated between NAFL and control samples and between NASH and control samples, respectively (Figure 1D; Data S2). Enrichment analysis showed that downregulated DEGs enriched in pathways associated with oxidative phosphorylation, spliceosome, thermogenesis, and proteasome (Figure 1E).

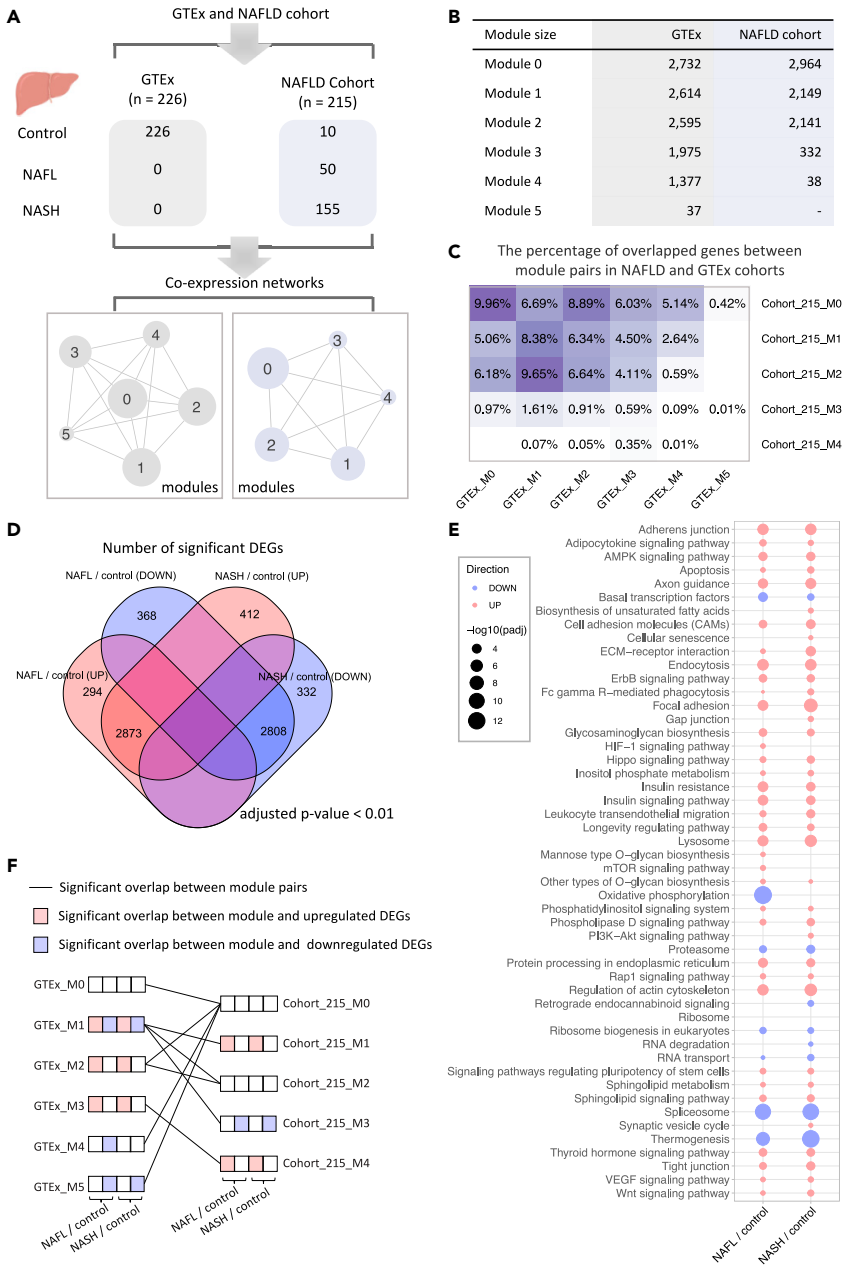


Figure 1. Sample information of studied cohorts and construction of co-expression networks

(A) Transcriptome data of liver tissue were obtained from GTEx, NAFLD cohort with 226 and 215 samples ranging from normal, NAFL, NASH, respectively. Spearman rank-order correlation coefficient analysis was applied to calculate the correlation between gene pairs after removing the lowly expressed genes (TPM < 1), and the Leiden algorithm was used to detect modules of significantly correlated genes. The label (number) of the module assigned by the algorithm.

(B) The numbers of genes consist of the individual module in each cohort.

(C) The heatmap shows the percentage of overlapped genes between module pairs in NAFLD and GTEx cohorts.

(D) The Venn diagram of differentially expressed genes (adjusted p value < 0.01) between patients with NAFLD and control samples.

(E) KEGG pathway analysis shows pathways that were significantly altered in patients with NAFLD. Up- and down-regulated pathways are shown in blue and red, respectively—only pathways with adjusted p value (padj) < 0.01 are presented (see also Data S3. The size of the bubble is scaled by $-\log_{10}$(padj) for each KEGG pathway term.

(F) Significant ($p < 0.05$, hypergeometric test) overlap between module pairs between GTEx and NAFLD cohorts and overlap between the module and dysregulated genes associated with NAFLD. GTEx, genotype-tissue expression; NAFLD, non-alcoholic fatty liver disease; NAFL, non-alcoholic fatty liver; NASH, non-alcoholic steatohepatitis; TPM, transcripts per kilobase million.


Reporter metabolites
Subsystems in iHepatocytes2322
Associated with dysregulated genes
N–acetylornithine[c]
L–glutamate 5–semialdehyde[m]
L–glutamate 5–semialdehyde[c]
L–4–hydroxyglutamate semialdehyde[m]
5–methylthioadenosine[c]
Arginine and proline metabolism
3–oxo–cis,cis–7,10–hexadecadienoyl–CoA[p]
3–oxo–cis,cis–5,8–tetradecadienoyl–CoA[p]
Beta oxidation of di–unsaturated fatty acids (n–6) (peroxisomal)
butyryl–CoA[m]
Beta oxidation of even–chain fatty acids (mitochondrial)
3–oxoeicosanoyl–CoA[p]
3–oxododecanoyl–CoA[p]
3–oxodocosanoyl–CoA[p]
3–oxodecanoyl–CoA[p]
Beta oxidation of even–chain fatty acids (peroxisomal)
lanosterol[c]
desmosterol[c]
Cholesterol biosynthesis 1 (Bloch pathway)
heparan sulfate, precursor 9[g]
heparan sulfate, precursor 8[g]
heparan sulfate, precursor 7[g]
heparan sulfate, precursor 6[g]
heparan sulfate, precursor 5[g]
heparan sulfate, precursor 4[g]
heparan sulfate, precursor 3[g]
heparan sulfate, precursor 2[g]
heparan sulfate, precursor 1[g]
Chondroitin / heparan sulfate biosynthesis
N–acetylmethionine[c]
Cysteine and methionine metabolism
oleoyl–CoA[c]
Fatty acid activation (cytosolic)
malonyl–CoA[m]
Fatty acid biosynthesis (even–chain)
THF[s]
THF[l]
pentaglutamyl–folate(THF)[s]
pentaglutamyl–folate(THF)[l]
pentaglutamyl–folate(DHF)[s]
pentaglutamyl–folate(DHF)[l]
molybdopterin[c]
hexaglutamyl–folate(THF)[s]
hexaglutamyl–folate(THF)[l]
hexaglutamyl–folate(DHF)[s]
hexaglutamyl–folate(DHF)[l]
heptaglutamyl–folate(THF)[s]
heptaglutamyl–folate(THF)[l]
heptaglutamyl–folate(DHF)[s]
heptaglutamyl–folate(DHF)[l]
dihydrofolate[s]
dihydrofolate[l]
10–formyl–THF[s]
10–formyl–THF[l]
10–formyl–THF–glu(7)[s]
10–formyl–THF–glu(7)[l]
10–formyl–THF–glu(6)[s]
10–formyl–THF–glu(6)[l]
10–formyl–THF–glu(5)[s]
10–formyl–THF–glu(5)[l]
Folate metabolism
tagatose–1,6–bisphosphate[c]
Galactose metabolism
acyl–CoA–LD–SM pool[c]
acyl–CoA–LD–PS pool[c]
acyl–CoA–LD–PI pool[c]
acyl–CoA–CL pool[m]
acyl–CoA–bile–PC pool[c]
1–acylglycerol–3P–LD–TG1 pool[c]
1–acylglycerol–3P–LD–SM pool[c]
1–acylglycerol–3P–LD–PS pool[c]
1–acylglycerol–3P–LD–PI pool[c]
1–acylglycerol–3P–LD–PE pool[c]
1–acylglycerol–3P–LD–PC pool[c]
1–acylglycerol–3P–CL pool[m]
1–acylglycerol–3P–bile–PC pool[c]
Glycerophospholipid metabolism
THF–hexaglutamate[c]
N–acetyl–L–alanine[c]
Glycine, serine and threonine metabolism
[protein]–N6–(lipoyl)lysine[c]
Lipoic acid metabolism
nm4masn[g]
dolichyl–diphosphate[c]
N–glycan metabolism
UMP[r]
UDP[s]
IMP[s]
IDP[s]
Nucleotide metabolism
ubiquinone[m]
ubiquinol[m]
ferrocytochrome C[m]
ferricytochrome C[m]
Oxidative phosphorylation
citrate[m]
Tricarboxylic acid cycle and glyoxylate/dicarboxylate metabolism
pyridoxine[c]
pyridoxine–phosphate[c]
pyridoxamine[c]
pyridoxamine–phosphate[c]
Vitamin B6 metabolism
−log10(p−value)
2 4 6 8 10
NAFL / control (up)
NASH / control (up)
NAFL / control (down)
NASH / control (down)

Figure 2. The most significant reporter metabolites between patients with NAFLD and control samples through the employment of *iHepatocytes2322*

Reporter metabolites were calculated for the up- and down-regulated genes for each comparison. Top30-ranked reporter metabolites and subsystems in *iHepatocytes2322* associated with up- and down-regulated genes in each comparison are presented, respectively. Color is proportional to the minus logarithm of the p value ($-\log_{10}$(p value)), see also Data S4.

We also examined the enrichment of those dysregulated DEGs associated with NAFLD in each co-expression module identified in GTEx and NAFLD cohort data. The results showed that module 1 and module 3 of the NAFLD cohort with 215 samples (cohort_215_M1 and cohort_215_M3) are significantly enriched (hypergeometric test p value ≈ 0) by upregulated and downregulated DEGs associated with NAFLD, respectively (Figure 1F). In particular, 58.3% of genes (1,253 of 2,149) in cohort_215_M1 and 92.2% of genes (306 of 332) in cohort_215_M3 are significantly upregulated and downregulated in NAFL vs control groups, respectively. Of all genes, 55.1% of genes in cohort1_215_M1 and 90.7% of genes (301 of 332) in cohort_215_M3 are significantly upregulated and downregulated in NASH vs control groups, respectively. Notably, we found both cohort_215_M1 and cohort_215_M3 are significantly overlapped (hypergeometric test p value = 2.17×10^{-14} and 3.11×10^{-8}) with module 1 in GTEx cohort (GTEx_M1), which were overrepresented by both upregulated and downregulated DEGs associated with NAFLD (Figure 1F). Interestingly, KEGG enrichment analysis of genes in those modules suggests that the significantly enriched pathways are consistent with the dysregulated pathways enriched by DEGs (Figures 1E and S2A). Moreover, we found cohort_215_M4 are significantly enriched by upregulated genes (15 and 21 of 38 in NAFL vs control and NASH vs control, respectively) and only significantly overlapped with GTEx_M3 (hypergeometric test p value = 4.11×10^{-13}). Taken together, co-expression network analysis identified modules of genes that are significantly perturbed in patients with NAFLD.

Altered metabolism in NAFLD patients

To further evaluate the detailed metabolic changes underlying NAFLD progression, we identified reporter metabolites (Patil and Nielsen, 2005), around which the most significant transcriptional changes occur, using differential expression data from NAFLD and network topology provided by *iHepatocytes2322* (Mardinoglu et al., 2014) (Figure S2B). Such reporter metabolites can thus be used to identify the key dysregulated regions of the metabolic network. A total of 321 metabolites were significantly (p value <0.05) associated with upregulated genes in either NAFL vs control or NASH vs control (Figures 2 and S2C; Data S4). Among these, the most significant reporter metabolites associated with upregulated genes in NAFL vs control were those involved in arginine and proline metabolism, glycerophospholipid metabolism, and nucleotide metabolism. The top reporter metabolites associated with upregulated genes between NASH and control samples were involved in beta-oxidation of fatty acids, cholesterol biosynthesis, and chondroitin/heparan sulfate biosynthesis. Chondroitin sulfate (CS) and heparan sulfate (HS) are the essential components of proteoglycans (PGs), which have been proposed as potential biomarkers for NASH diagnosis and staging of NAFLD by integrative analysis of transcriptomic data obtained from patients with NAFLD and GEM (Mardinoglu et al., 2014). The analyses from the current investigation are utterly consistent with the previous study. In addition, we observed 215 metabolites that were significantly associated with downregulated genes in NAFLD, involving in folate metabolism and oxidative phosphorylation (Figures 2 and S2D; Data S4).

Validation of perturbed modules in two independent NAFLD cohorts

To validate whether modules related to significant transcriptomics and metabolic changes in patients with NAFLD can truly reflect the perturbations in a disease-specific manner, we analyzed GCNs generated using liver-biopsy transcriptomics data sets from two independent NAFLD cohorts with 75 and 58 samples, respectively (Azzu et al., 2021; Hoang et al., 2019). To avoid repeated IDs, we assigned cohort 1 to the studied NAFLD cohort, and 2 and 3 for NAFLD cohorts for validation in the downstream analysis. By the same method of constructing GCN for the first NAFLD cohort, we identified four and eight modules of genes in NAFLD cohort 2 and 3, respectively (Figures S3A and S3B). To explore module similarity among NAFLD cohorts, we calculated the Jaccard index between each pair of modules from different NAFLD cohorts and performed hypergeometric test to evaluate the significance of the observed overlap in gene members (Figures 3A, 3B, and 3C; Data S1). To begin with, we tested the modules between NAFLD cohort 1 and cohort 2. The results showed that genes in cohort1_215_M4 were only significantly overlapped (29 of 38; Jaccard index = 0.388; hypergeometric test p value = 1.66×10^{-69}, Figures 3A and 3D) with genes in module 3 of NAFLD cohort 2 with 75 samples (cohort2_75_M3). We next tested the module pairs in NAFLD

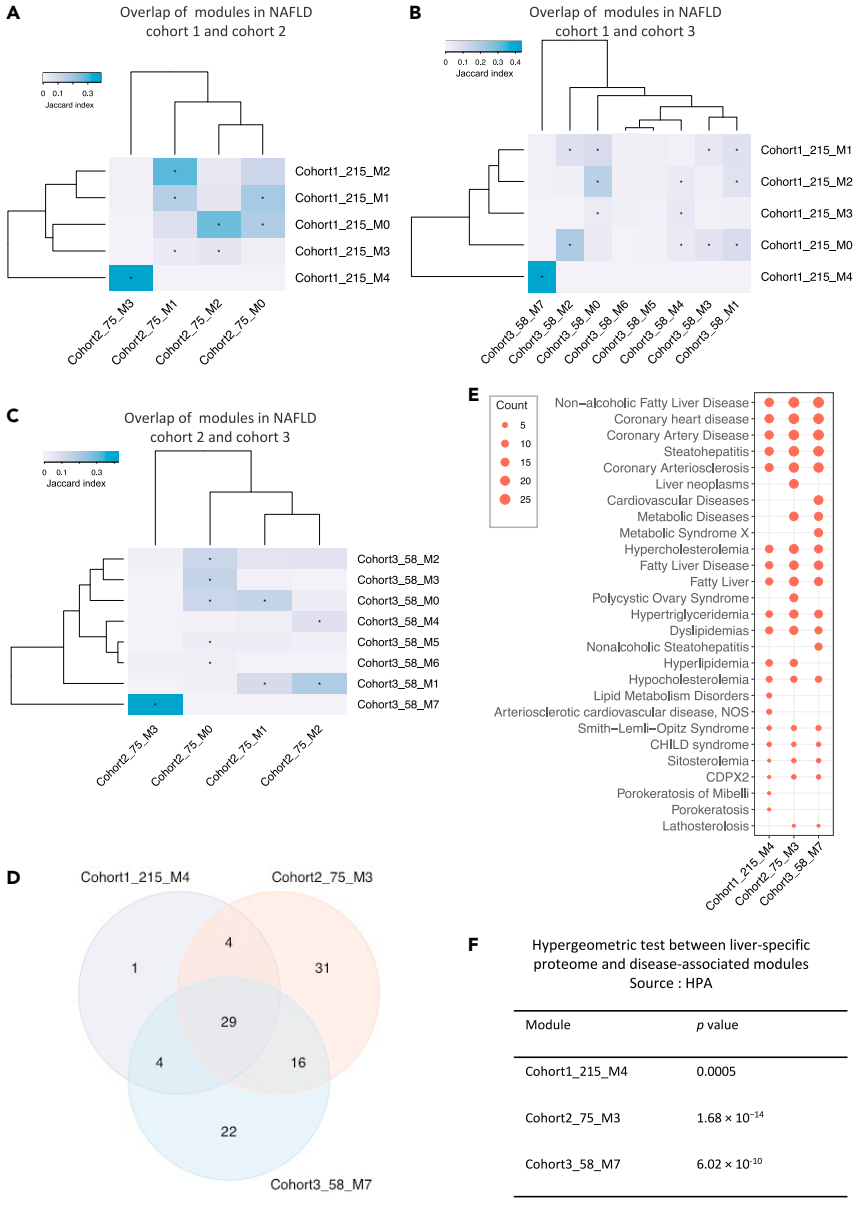


Figure 3. Validation of disease-related modules using two independent NAFLD cohorts

(A–C) Hierarchical clustering of Jaccard Index between module pairs from NAFLD cohort 1 and 2; NAFLD cohort 1 and 3; NAFLD cohort 2 and 3. Color scales representing the range of the Jaccard index. Asterisk indicates the statistical significance of the overlap between gene members in any two modules from the different cohorts.

(D) Venn diagram shows numbers of genes overlapped between cohort1_215_M4, cohort2_75_M3, and cohort3_58_M7.

(E) Dot-plot heatmap shows top 20 significantly ("q-value FDR B&H" < 0.05) enriched diseases by genes in each module (cohort1_215_M4, cohort2_75_M3, and cohort3_58_M7). The size of each dot is proportional to the number of genes enriched in each disease term.

(F) The table shows the results from a hypergeometric test between liver-specific proteome (HPA) and disease-associated modules in NAFLD cohorts, the overlap with p value less than 0.05 was considered as significant.

cohort 1 and cohort 3, and found cohort1_215_M4 also shared 29 genes with module 7 of NAFLD cohort 3 with 58 samples (cohort_3_58_M7) (Jaccard index = 0.434; hypergeometric test p value = 3.73×10^{-71}, Figures 3B and 3D). Interestingly, the genes in cohort2_75_M3 were significantly overlapped (45 of 80; Jaccard index = 0.425; hypergeometric test p value = 1.82×10^{-86}, Figures 3C and 3D) with genes in cohort3_58_M7 as well.

To validate if those conserved modules in NAFLD cohorts have a similar expression pattern in normal liver tissue, we subsequently assessed the module similarity and overlap between any two modules between the GTEx cohort and NAFLD cohorts. Hierarchical clustering of the Jaccard index between module pairs showed that a distinct cluster consisting of cohort1_215_M4, cohort2_75_M3, and cohort3_58_M7 was only significantly over-represented by module 3 in GTEx cohort (GTEx_M3), which contains 1,975 genes (Figures 1B and S3C). We found that more than 70% genes in cohort1_215_M4 (27 of 38; hypergeometric test p value = 4.11×10^{-13}), cohort2_75_M3 (57 of 80; hypergeometric test p value = 4.97×10^{-27}), and cohort3_58_M7 (50 of 71; hypergeometric test p value = 2.92×10^{-24}) were included by module GTEx_M3 (Figure 4A).

For a systematic evaluation on biological functions related to the modules, we quantified the statistical significance of enrichment of genes with the association in disease-related gene sets obtained from DisGeNET database (https://www.disgenet.org/) (Pinero et al., 2020), liver-specific proteome in Human Protein Atlas (HPA) database (http://www.proteinatlas.org/) (Uhlen et al., 2015), and KEGG pathway gene sets. We found that genes in cohort1_215_M4, cohort2_75_M3, and cohort3_58_M7 were significantly over-represented by multiple liver disease-related gene sets, including NAFLD and steatohepatitis (Figure 3E; Data S5). Interestingly, these three modules were also significantly enriched in coronary heart disease, coronary artery disease, and coronary arteriosclerosis. We further evaluated the overlap of genes in each of three disease-associated modules with 936 liver-specific genes defined by HPA (Fagerberg et al., 2014; Uhlen et al., 2015; Yu et al., 2015). The results showed that genes in cohort1_215_M4 (10 out of 38; hypergeometric test p value = 0.0005), cohort2_75_M3 (30 out of 80; hypergeometric test p value = 1.68×10^{-14}), and cohort3_58_M7 (22 out of 71; hypergeometric test p value = 6.02×10^{-10}) are highly enriched with liver-specific genes (Figure 3F). In addition, we observed that genes in GTEx_M3, which shows high module similarity with those three modules identified in diseases cohorts, are significantly enriched in the peroxisome, branched-chain amino acids (BCAAs; valine, leucine, and isoleucine) degradation, and fatty acid degradation (Figure 4B). However, steroid biosynthesis and terpenoid backbone biosynthesis were the most significantly enriched pathways in all the three modules of disease cohorts. Moreover, fatty acid biosynthesis, citrate cycle (TCA cycle), and insulin signaling pathway were only significantly enriched in the modules of disease cohort(s).

Topological features of genes in NAFLD-associated modules

The analysis of topological properties can provide important information about hub genes or other influential genes in the module. To understand the interplay of genes in the module, we then obtained several key network properties using the "NetworkAnalyzer" in Cytoscape to analyze the disease-associated modules. In our workflow, we used degree and closeness centrality to evaluate the importance of nodes in a module. In an undirected network, the degree of a node is the number of edges linked to this node and a node with a high degree has been considered as functionally significant (Doncheva et al., 2012). Genes with high closeness centrality are considered as a controlling point of molecular communication (Miryala and Ramaiah, 2019).

Topological analysis showed that gene *FDFT1* has the highest degree in both cohort1_215_M4 (26 edges) and cohort2_75_M3 (41 edges), whereas gene *MVD* has the highest degree in cohort3_58_M7 (Figures 4C–4E, S4, and S5A; Data S6). The other genes with high connectivity are *HMGCS1*, *DHCR7*, and *ACSS2* (23 edges, respectively), and *ACAT2* (22 edges) in cohort1_215_M4; *MVD* (40 edges), *MSMO1* and *SQLE* (37 edges, respectively), and *IDI1* and *NSDHL* (36 edges, respectively) in cohort2_75_M3; *IDI1*(30 edges), *LSS*, *CYP51A1*, *FDFT1*, and *SQLE* (29 edges, respectively) in cohort3_58_M7. Interestingly, we observed a highly positive correlation between the degree of 33 genes overlapped in cohort1_215_M4 and cohort2_75_M3 (Spearman's correlation = 0.58, p = 0.00036, Figure 4C), which indicates that those two modules have a similar topological structure. Similarly, a highly positive correlation between degree of 33 genes shared by cohort1_215_M4 and cohort3_58_M7 (Spearman's correlation = 0.73; $p = 1.7 \times 10^{-6}$, Figure 4D) and that of degree of 45 genes shared by cohort2_75_M3 and cohort3_58_M7 (Spearman's correlation = 0.81; $p = 1.7 \times 10^{-11}$, Figure 4E) were also observed. Moreover, the top five genes with the highest closeness centrality in cohort1_215_M4, cohort2_75_M3, and cohort3_58_M7 are also highly conserved (Figure S5B). We also observed a strong correlation between closeness centrality of shared genes in any disease-associated module pairs of NAFLD cohorts (Figure S5C).

Validation of topological features in an HCC cohort

Given that NAFLD has emerged as the fastest-growing cause of HCC (Huang et al., 2021; Ray, 2018), we next investigated whether the expression of the genes in NAFLD disease-associated modules, especially genes

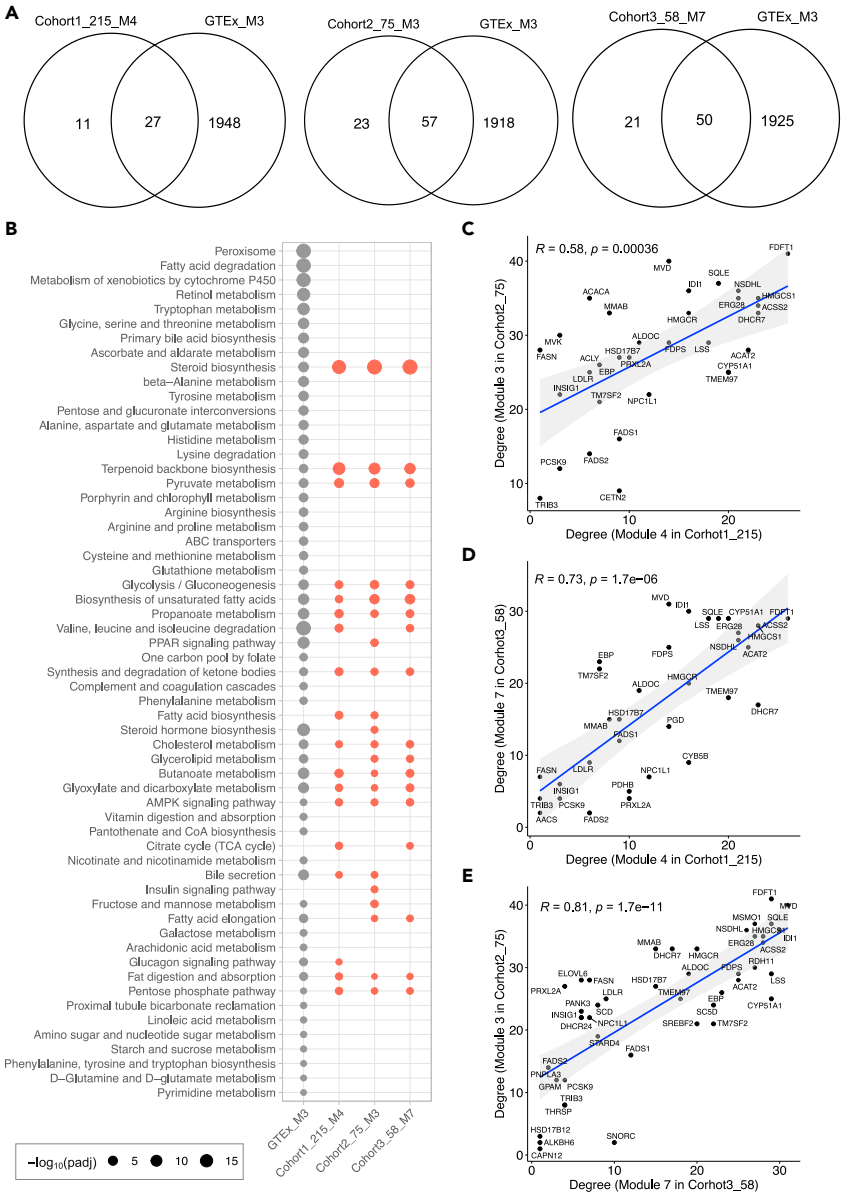


Figure 4. Functional enrichment and topological structure analyses of disease-associated modules

(A) Venn diagram shows numbers of genes overlapped between GTEx_M3 cohort and cohort1_215_M4 (hypergeometric test p value = 4.11×10^{-13}), cohort2_75_M3 (hypergeometric test p value = 4.97×10^{-27}), and cohort3_58_M7 (hypergeometric test p value = 2.92×10^{-24}), respectively.

(B) Dot-plot heatmaps are showing KEGG pathways enriched in different modules. The color differences of dots indicate the studied cohort (GTEx or NAFLD) in which the module detected. The size of each dot is proportional to the significance ($-\log_{10}$(padj); padj represents "q-value FDR B&H" with value <0.05) of enrichment for each KEGG pathway term.

(C–E) Correlation between degrees among disease-associated modules from different cohorts. The correlation was evaluated by Spearman correlation coefficients.

with high-connectivity, is predictive of patients with HCC using the Liver Hepatocellular Carcinoma dataset (https://portal.gdc.cancer.gov/projects/TCGA-LIHC) (Figure S6A). The expression of 19 genes in cohort1_215_M4, 39 genes in cohort2_75_M3 and 40 genes in cohort3_58_M7 are significantly (log rank p value < 0.05) associated with the survival of patients, respectively (Figure S6B; Data S7). Among these, the high expression of 19 genes in cohort1_215_M4, 28 genes in cohort2_75_M3, and 31 genes in cohort3_58_M7 are significantly associated with an unfavourable survival of patients. For example, the high expression of *FDFT1* (log

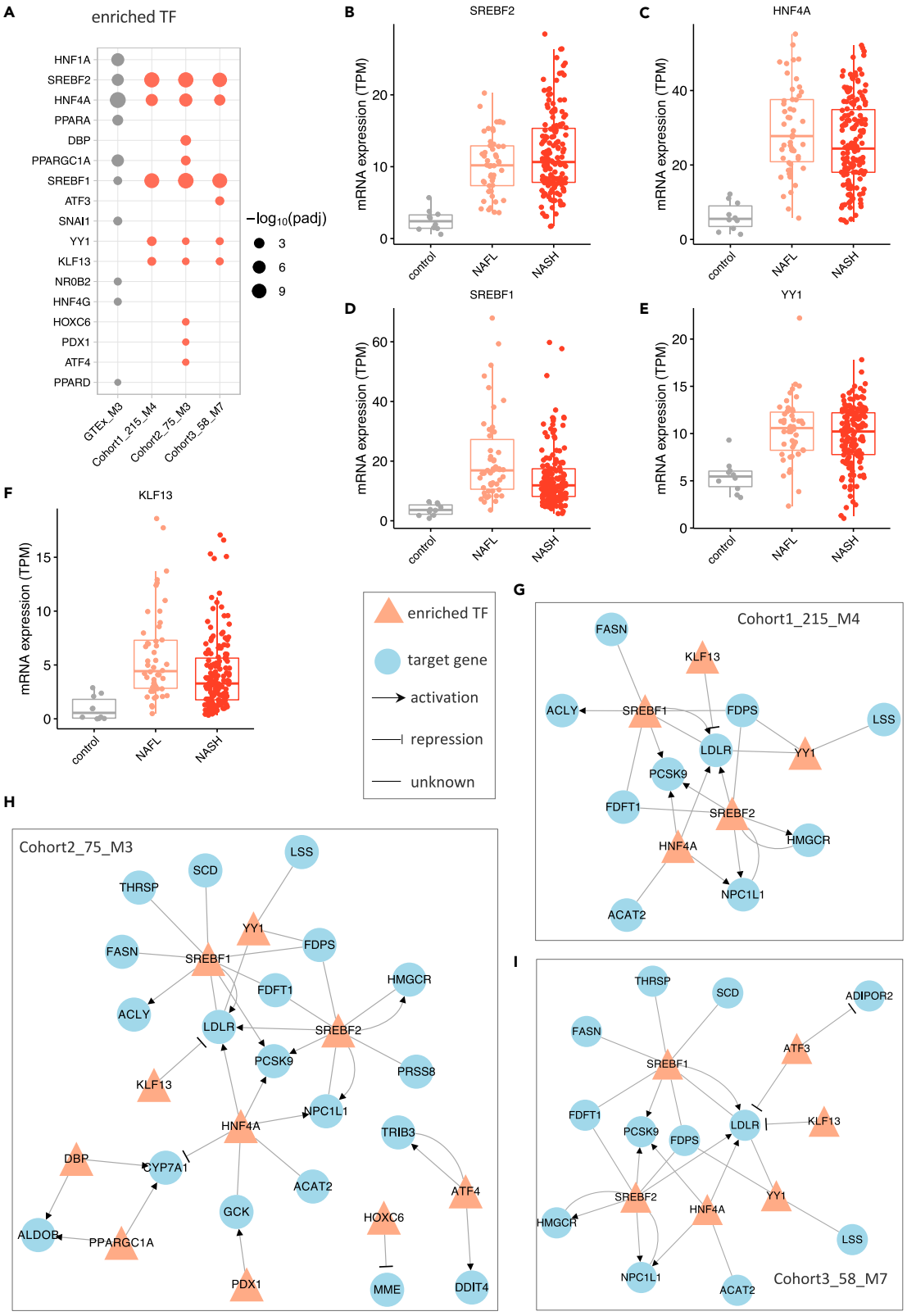

A
enriched TF
HNF1A
SREBF2
HNF4A
PPARA
DBP
PPARGC1A
SREBF1
ATF3
SNAI1
YY1
KLF13
NR0B2
HNF4G
HOXC6
PDX1
ATF4
PPARD
GTEx_M3
Cohort1_215_M4
Cohort2_75_M3
Cohort3_58_M7
−log10(padj)
3
6
9
B
SREBF2
C
HNF4A
D
SREBF1
E
YY1
F
KLF13
mRNA expression (TPM)
control
NAFL
NASH
enriched TF
target gene
activation
repression
unknown
G
Cohort1_215_M4
FASN
KLF13
ACLY
SREBF1
FDPS
LSS
LDLR
YY1
PCSK9
FDFT1
SREBF2
HMGCR
HNF4A
NPC1L1
ACAT2
H
Cohort2_75_M3
LSS
SCD
THRSP
YY1
FASN
SREBF1
FDPS
HMGCR
FDFT1
ACLY
LDLR
SREBF2
PCSK9
PRSS8
KLF13
NPC1L1
HNF4A
TRIB3
DBP
CYP7A1
ACAT2
ATF4
GCK
HOXC6
ALDOB
PPARGC1A
PDX1
MME
DDIT4
I
Cohort3_58_M7
THRSP
SCD
ADIPOR2
FASN
ATF3
SREBF1
FDFT1
PCSK9
FDPS
LDLR
KLF13
SREBF2
HNF4A
YY1
HMGCR
LSS
NPC1L1
ACAT2

Figure 5. Regulatory relationship between enriched transcription factors and associated target genes in disease-associated modules
(A) enriched transcription factors in GTEx_M3, cohort1_215_M4, cohort2_75_M3, and cohort3_58_M7.
(B–F) mRNA hepatic expression of the enriched transcription factors including SREBF2, HNF4A, SREBF1, YY1, and KLF13.
(G–I) the regulatory network between enriched transcription factors and associated target genes in cohort1_215_M4, cohort2_75_M3, and cohort3_58_M7, repectively. The regulation between transcription factor and its target was retrieved from the TRRUST database.

rank p value = 6.54×10^{-4}) with the highest connectivity in both cohort1_215_M4 and cohort2_75_M3 and *MVD* (log rank p value = 1.26×10^{-3}) with the highest connectivity in cohort3_58_M7 are significantly associated with poor patient outcome (Figures S6C and S6D). In addition, some of these genes have already been described as associated with NAFLD associated HCC (NAFLD-HCC). For instance, the high expression of *SQLE*, a second rate-limiting enzyme involved in *de novo* cholesterol synthesis with relatively high connectivity in disease-associated modules (Figures S4 and S5A), was predictive of unfavourable survival of HCC patients (log rank p value = 7.39×10^{-4}; Figure S7). Indeed, recent studies have demonstrated that *SQLE* acts as an independent prognostic factor in patients with NAFLD-HCC, and *SQLE* inhibition suppressed NAFLD-HCC growth *in vitro* and *in vivo* (Liu et al., 2018a; Ray, 2018).

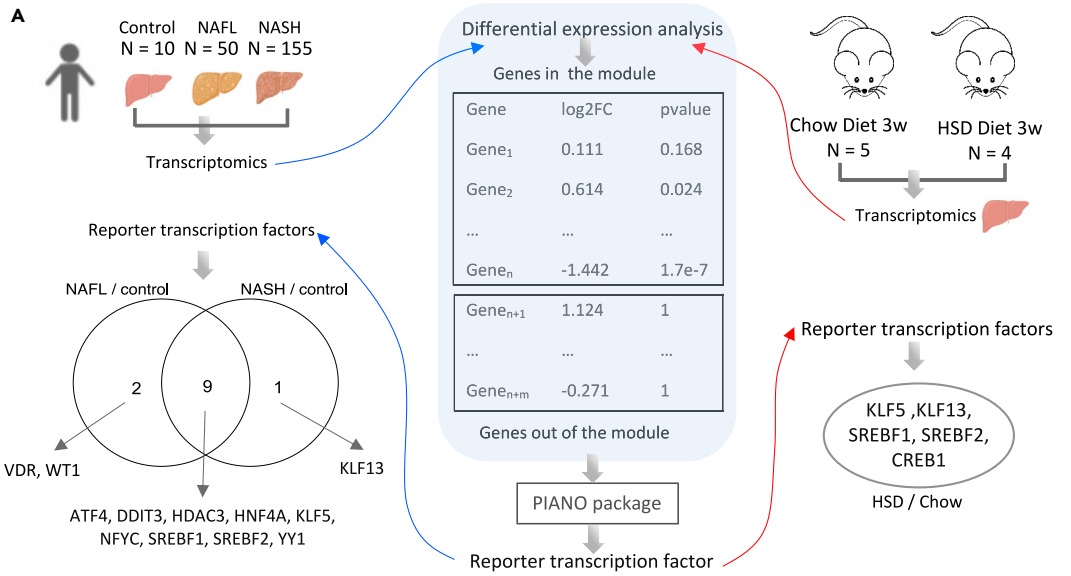
Identification of TFs that regulate the NAFLD-associated modules

To investigate the transcriptional regulation in maintaining homeostasis and alterations in the disease state, we performed TF enrichment analysis (hypergeometric test) using the genes from the disease-associated modules and module 3 in GTEx, which shows high similarity to disease-associated modules (Figure 5; Data S8), based on TRRUST database (Han et al., 2018). Our results indicated that *HNF4A*, *HNF1A*, *PPARGC1A*, *SREBF2*, and *PPARA* are the most significantly enriched TFs in GTEx_M3 (Figure 5A). We observed that *HNF4A*, *SREBF1*, *SREBF2*, *YY1*, and *KLF13* are significantly enriched TFs in all three disease-associated modules. We also found significant upregulation of hepatic expression of *SREBF1*, *SREBF2*, *HNF4A*, and *KLF13* in NAFL and NASH compared to the control group (adjusted p value < 0.05, Figures 5B–5F; Data S2).

We then constructed the regulatory networks for the enriched TFs and associated targets in each of the modules (Figures 5G–5I and S8). We observed that *HNF4A*, an important transcriptional factor mainly expressed in the liver, regulates the expression of genes involved in lipid metabolism and fatty acid oxidation, including cholesterol/triglyceride transporter (e.g., *ABCG8, ABCG5* and *MTTP*), oxidoreductase (e.g., *AKR1C4*, *CYP2D6,* and *CYP2B6*) in the regulatory network of GTEx_M3 (Figure S8). As known, *SREBF1* and *SREBF2* regulate the expression of genes associated with *de novo* lipogenesis (DNL) (e.g., *FASN, SCD, ACACB*), synthesis and cellular uptake of cholesterol (e.g., *HMGCR*, *FDFT1*, *NPC1L*), respectively. Moreover, *PPARA* regulates the expression of genes involved in peroxisomal and mitochondrial β-oxidation, including *ACSL1*, *CPT1A*, *CYP1A1*, and *ACOX1*. Apolipoprotein C3 (*APOC3*), a central regulator of plasma triglyceride levels by inhibiting the removal of remnants of triglyceride-rich lipoproteins, is the most highly regulated gene by *HNF4A*, *NR0B2*, *PPARA,* and *PPARGC1A* in the regulatory network of GTEx_M3. Interestingly, low-density lipoprotein receptor (*LDLR*, a key receptor that is internalized by endocytosis) is the most highly regulated genes in the disease-associated modules (Figures 5G–5I) by *SREBF1*, *SREBF2*, *HNF4A*, *YY1,* and *KLF13*. This indicates that highly co-expressed genes involved in cholesterol metabolism in disease-associated modules are essential compared with the other endocytosis-related genes that are co-expressed in other modules in the same cohort. In addition to the well-established regulation of *LDLR* activation by SREBFs and *HNF4A*, *YY1* and *KLF13*, two specific TFs regulating the disease-associated modules, also showed a regulatory role in the transcriptional regulation of *LDLR*. Taken together, the complicated regulation of *LDLR* in the disease-associated modules rather than endocytosis in normal liver tissue might play an essential role in the dysregulation of lipid metabolism underlying the NAFLD pathogenesis.

Validation of TFs in a mouse NAFLD model

Next, we generated liver transcriptomics data from a mouse NAFLD model fed by HSD and performed reporter TF analysis (Huang et al., 2017; Liu et al., 2019; Oliveira et al., 2008) by integrating with the same network of TF-target from TRRUST database (Han et al., 2018). We validated the TFs that are enriched in disease-associated modules (Figure 6A). The reporter TF algorithm was used to calculate the statistically significant expression changes of gene sets controlled by TFs. To study the regulation of module of genes using this method, we first examined the reporter TFs that are significantly associated with the upregulated and downregulated genes in NAFL vs control and NASH vs control, respectively (Data S9). The analysis identified 12 reporter TFs of genes in cohort1_215_M4, of which 9 were associated with upregulated genes


A
Control NAFL NASH
N = 10 N = 50 N = 155
Transcriptomics
Differential expression analysis
Genes in the module
Gene log2FC pvalue
Gene1 0.111 0.168
Gene2 0.614 0.024
... ...
Genen -1.442 1.7e-7
Genen+1 1.124 1
... ...
Genen+m -0.271 1
Genes out of the module
PIANO package
Reporter transcription factor
Chow Diet 3w N = 5
HSD Diet 3w N = 4
Transcriptomics
Reporter transcription factors
NAFL / control
NASH / control
2
9
1
VDR, WT1
KLF13
ATF4, DDIT3, HDAC3, HNF4A, KLF5, NFYC, SREBF1, SREBF2, YY1
Reporter transcription factors
KLF5 ,KLF13, SREBF1, SREBF2, CREB1
HSD / Chow

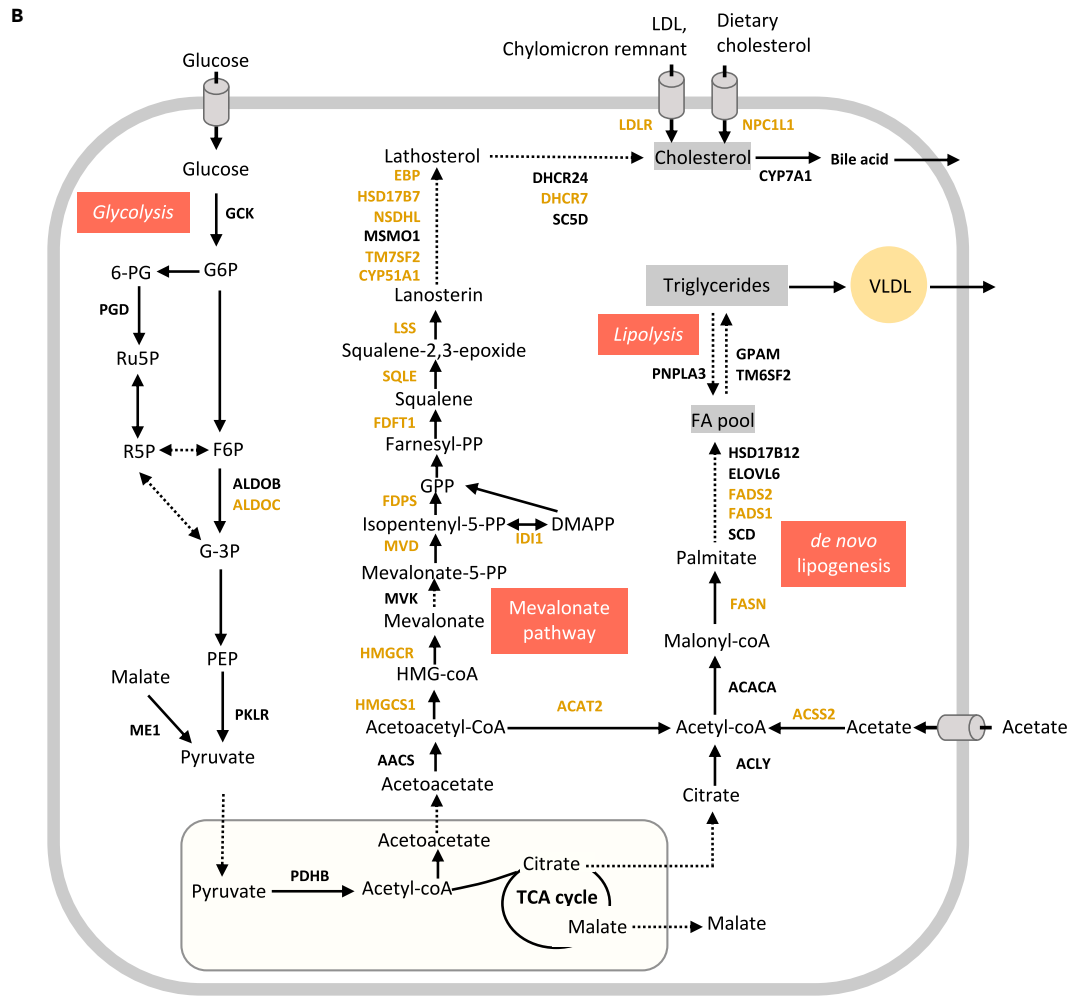

B
Glucose
Glucose
Glycolysis
GCK
6-PG
G6P
PGD
Ru5P
R5P
F6P
ALDOB
ALDOC
G-3P
PEP
Malate
ME1
PKLR
Pyruvate
LDL, Chylomicron remnant
Dietary cholesterol
LDLR
NPC1L1
Lathosterol
Cholesterol
Bile acid
CYP7A1
EBP
HSD17B7
NSDHL
MSMO1
TM7SF2
CYP51A1
DHCR24
DHCR7
SC5D
Lanosterin
LSS
Squalene-2,3-epoxide
SQLE
Squalene
FDFT1
Farnesyl-PP
FDPS
GPP
DMAPP
IDI1
Isopentenyl-5-PP
MVD
Mevalonate-5-PP
MVK
Mevalonate
Mevalonate pathway
HMGCR
HMG-coA
HMGCS1
ACAT2
Acetoacetyl-CoA
AACS
Acetoacetate
Triglycerides
VLDL
Lipolysis
PNPLA3
GPAM
TM6SF2
FA pool
HSD17B12
ELOVL6
FADS2
FADS1
SCD
Palmitate
de novo lipogenesis
FASN
Malonyl-coA
ACACA
Acetyl-coA
ACSS2
Acetate
Acetate
ACLY
Citrate
Acetoacetate
Pyruvate
PDHB
Acetyl-coA
Citrate
TCA cycle
Malate
Malate

Figure 6. Validation of transcription factors in a mouse NAFLD model

(A) Reporter transcription factor analysis was used to validate transcription factors identified in disease-associated modules using transcriptomics data of NAFLD cohort 1 and newly generated from a mouse NAFLD model.

(B) Conserved disease-associated modules revealed the dysregulation in the mevalonate pathway, *de novo* lipogenesis, glycolysis, and lipolysis.

in NAFLD, including, *ATF4*, *DDIT3*, *HDAC3*, *HNF4A*, *KLF5*, *NFYC*, *SREBF1*, *SREBF2*, and *YY1*. 2 reporter TFs (*VDR* and *WT1*) are associated explicitly with downregulated genes in cohort_215_M4 between patients with NAFL and control samples. *KLF13* was significantly associated with upregulated genes between NASH and control samples. Among these reporter TFs, five of them (*HNF4A*, *SREBF1*, *SREBF2*, *YY1*, and *KLF13*) were also identified by hypergeometric test for cohort1_215_M4 (Figure 5A). Between mice fed by HSD and chow diet (Figure 6A; Data S10), 5 reporter TFs (including *KLF5*, *KLF13*, *SREBF1*, *SREBF2*, and *CREB1*) were identified, showing significant association with upregulation of genes using corresponding human orthologs. Taken together, our analysis validated *SREBF1*, *SREBF2*, and *KLF13* as TFs that regulate the hepatic expression of genes in cohort1_215_M4.

Hepatic co-expression networks reflect dysregulated cholesterol homeostasis and de novo lipogenesis in the NAFLD cohorts

In our network analysis, we found conserved disease-associated modules across three independent NAFLD cohorts and more than 70% of the genes involved in the modules are associated with metabolic functions. Most of the metabolic genes in this consensus module are associated with cholesterol metabolism. For instance, 13 genes, namely *HMGCS1*, *HMGCR*, *MVD*, *IDI1*, *FDPS*, *FDFT1*, *SQLE*, *LSS*, *CYP51A1*, *TM7SF2*, *NSDHL*, *HSD17B7*, *EBP*, and *DHCR7*, which are shared among the three disease-associated modules from different cohorts, and 5 genes, namely *AACS*, *MVK*, *MSMO1*, *DHCR24*, and *SC5D* which are included in at least one of the disease-associated modules from different cohorts, are involved in the endogenous synthesis of cholesterol (Figure 6B). *LDLR* and *NPC1L1*, responsible for the uptake of cholesterol, are also found in the disease-associated modules from all three cohorts. In addition, several genes, namely *ACLY*, *ACSS2*, *ACACA*, *FASN*, *SCD*, *FADS1*, *FADS2*, *ELOVL6*, *HSD16B12*, *GPAM*, *PNPLA3*, and *TM6SF2*, which are involved in *de novo* lipogenesis and lipolysis, are also included in the disease-related modules in at least one of the cohorts. Finally, genes encoding glycolytic enzymes, such as *GCK*, *PGD*, *ALDOB*, *ALDOC*, *PKLR*, *ME1*, and *PDHB*, are also found in the disease-related modules. In summary, co-expression network analysis revealed a strong connection between the disease-associated clusters with the cholesterol metabolism, *de novo* lipogenesis and glycolysis in the liver and suggests their potential roles in the development of NAFLD.

DISCUSSION

Here, we applied a systems biology approach on human liver transcriptomics data to elucidate the dysregulated biological processes involved in NAFLD and identified potential regulators via integrating with a transcriptional regulatory network. Our analysis identified highly conserved disease-associated gene modules across three different NAFLD cohorts. These modules are specific to the disease networks, and we could not find such modules in the network generated from normal subjects (GTEx cohort). Therefore, these gene modules could play a critical role in the development of NAFLD indicating their importance to the mechanism of the disease. Interestingly, we found the majority of the genes (~70%) in these disease-associated modules identified in the NAFLD cohort are included in the big gene module 3 of the GTEx cohort, which has 1,975 genes, suggesting that the disease-associated module and its related biological functions are co-regulated with a large gene group in normal subjects and dysregulated with the progression of NAFLD.

In addition, we showed that enriched TFs that regulate the disease-associated modules, which can facilitate our understanding of the regulatory mechanism of these perturbed biological processes. Transcription regulatory networks analysis indicated that *SREBF1*, *SREBF2*, *HNF4A*, *YY1*, and *KLF13* are the most prominent regulators of gene expression in disease-associated modules, 3 of which (*SREBF1*, *SREBF2*, and *KLF13*) were validated using the transcriptomics data generated from a mouse NAFLD model. Notably, *KLF13* is reporter TF, specific to this disease-associated module but not for the module from the normal subjects, suggesting their potential role in the development of NAFLD (Ericsson et al., 1999; Natesampillai et al., 2006). It has been shown that selective overexpression of *YY1* results in massive triglyceride accumulation and moderate insulin resistance in mice fed with HFD (Lu et al., 2014), and it may be a promising target for fatty liver diseases (Wu et al., 2017). We also show that *LDLR* is a central target gene regulated

by the enriched TFs in this disease-associated module. It has been demonstrated that multiple mechanisms are involved in protecting against excessive cholesterol accumulation in the liver (Goldstein and Brown, 2009; Natesampillai et al., 2006). *LDLR*-mediated endocytosis contributes to this process by removing approximately 70% circulating cholesterol-enriched LDL and providing feedback transcriptional regulation of cholesterol synthesis through SREBFs (Goldstein and Brown, 2009).

Our systematic analyses also highlight the significant reporter metabolites involved in CS and HS biosynthesis, glycerophospholipid metabolism, folate metabolism, and oxidative phosphorylation. Such metabolites are consistent with the findings of previous studies and could be targeted for discovery of potential biomarkers in diagnosis of NAFLD. We also find that most of the genes involved in the disease-associated module are involved in metabolic pathways such as cholesterol metabolism, DNL, and glycolysis.

The liver plays a central role in cholesterol homeostasis, and growing evidence has shown that excess hepatic cholesterol and its associated hepatic lipotoxicity is a predominant factor in the development of human NAFLD (Ioannou, 2016; Min et al., 2012). Abundant hepatic free cholesterol stimulates Kupffer cells and hepatic stellate cells (HSCs), which are key mediators of fibrosis and inflammation, as well as mitochondrial dysfunction, and thus reflects the severity of disease (Musso et al., 2013). Notably, differential expression (DE) analysis pointed out significant upregulation of critical genes (adjusted p value < 0.05) in these cholesterol-related pathways in the disease-related module, including *HMGCR* (the principal rate-limiting enzyme in mevalonate pathway), *NPC1L1* (a major gene in intestinal cholesterol absorption) in NASH compared to control group (Data S2). We find scavenger receptor class B type I (*SCARB1*), which mediate the uptake of HDL cholesterol directly, is significantly increased in NAFLD patients compared to the control group. This suggests that upregulation of pathways in both synthesis and absorption of cholesterol may associate with the increased hepatic cholesterol (Ioannou, 2016), as well as increased bile acids in NAFLD patients (Jiao et al., 2018).

Additionally, recent studies have shown that high dietary cholesterol in the mice model is the causative factor for the progression of steatohepatitis to fibrosis and drives NAFLD associated HCC (Liu et al., 2018a; Shen et al., 2020; Zhang et al., 2021). We, therefore, investigated whether the disease-associated modules are predictive of patient outcome using the liver cancer data set. The results from Kaplan–Meier analysis show that the high expression of ~41% genes (12 of 29 genes identified in all three disease-associated modules) (Figures S6 and S7) are significantly associated with poor survival of patients, for example, *FDFT1*, *MVD*, *DHCR7*, *SQLE*, and *MVD* with high connectivity in these modules. Liu et al. demonstrated that targeting *SQLE* can efficiently inhibit the NAFLD-HCC in cellular and animal models (Liu et al., 2018a). Considering the characteristics of the co-expression mechanism among genes with similar functions, this integrative network analysis reveals detailed molecules involved in the cholesterol metabolism and thus proposes more potential therapeutical targets of effective treatment for preventing NAFLD-to-HCC progression.

Moreover, in the disease-associated module, we also find genes associated with DNL. Generally, it is believed that the triglyceride accumulation in the liver of NAFLD patients is caused by elevation of both DNL and fat uptake (Donnelly et al., 2005; Perdomo et al., 2019). However, we do not find any genes related to fat uptake in these disease-associated modules. In fact, *CD36*, the key free fatty acid transporter, is not significantly changed between NAFLD and the control group (Data S2). The hepatic expression of *FABP5*, another critical transporter for fat, is significantly lower in the patients than the control group. In addition, most of DNL related genes are significantly upregulated in the NAFL and NASH patients compared with the control group. Taken together, these findings suggest that the DNL, rather than free fatty acid uptake, is the source of triglyceride accumulation in NAFLD patients.

Finally, we identified a few key enzymes involved in glycolysis and insulin signaling pathway which are included in the disease-associated module. For instance, *INSIG*, a key player in the insulin signaling pathway is included in the disease-associated modules among all three cohorts. A recent study has reported that *INSIG* is a central regulator in a negative feedback loop ensuring the balance of lipid desaturation and cholesterol composition and loss of *INSIG1* improves liver damage and would healing NASH progression (Azzu et al., 2021). In addition, *GCK*, which is a kinase specific to glucose, is involved in the module of cohort 2, and it is significantly upregulated in the NAFLD patients compared with the normal subjects (Data S2). In our previous study, we have reported that the *GCK* up-regulation is associated

with elevated insulin resistance in patients and suggested an increased in influx from dietary glucose (Lee et al., 2016). Moreover, we also observed that *TKFC* included in disease-associated module and up-regulated in NAFLD patients. It has been reported that increased dietary fructose uptake could cause NAFLD in both mouse and human patients (Jensen et al., 2018; Loomba et al., 2021). Therefore, these results highlight the association between NAFLD and insulin resistance and suggest the potentially important contribution of dietary glucose, fructose, and sucrose to development of the disease.

In summary, unlike previous studies with the limitation of a few human NAFLD transcriptome data or focusing on individual genes influencing NAFLD progression, our network-driven approach reveals a highly conserved disease-associated gene module across three heterogeneous cohorts including patients with various degrees of NAFLD. In addition, our results highlight the predominant role of key transcriptional regulators, including *SREBF2*, *HNF4A*, *SREBF1*, *YY1*, and *KLF13* that are associated with lipid and cholesterol metabolism. Our integrative study enabled a comprehensive view of the molecular processes and key drivers associated with NAFLD, which provide molecular candidates in dysregulated pathways for developing effective therapies.

Limitations of the study

Although we have identified key transcriptional factors and validated their roles in dietary mice model, we did not validate them in disease models. Therefore, future studies *in vitro/in vivo* disease models will be required to further investigate the potential pathogenic roles of the identified TFs in NAFLD.

STAR★ METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
- METHOD DETAILS
  - Transcriptomics data from mouse model
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Data retrieving and pre-processing
  - Construction of co-expression network and analysis
  - Functional annotation of modules in co-expression network of cohort
  - Reporter metabolite and reporter transcription factor analyses
  - TCGA data process and survival analysis

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2021.103222.

ACKNOWLEDGMENTS

A.M. and H.Y. acknowledge support from the PoLiMeR Innovative Training Network (Marie Skłodowska-Curie Grant Agreement No. 812616) which has received funding from the European Union's Horizon 2020 research and innovation program. The authors would like to acknowledge financially support from ScandiBio Therapeutics and Knut and Alice Wallenberg Foundation (No. 72110). The computations were performed on resources provided by SNIC through Uppsala Multidisciplinary Center for Advanced Computational Science (UPPMAX) under Project sllstore2017024. We would like to thank Dr. Theo Portlock for the comments on the manuscript.

AUTHOR CONTRIBUTIONS

Conceptualization ,A.M., C.Z., and H.Y.; methodology, software, and formal analysis, H.Y., M.A., and M.Y.; investigation, H.Y., M.A., M.Y., and C.Z.; writing-original draft, H.Y.; writing - review & editing, H.Y., M.A., M.Y., X.L., K.S., H.T., J.N., M.U., J.B., C.Z., and A.M.; funding acquisition, A.M.; supervision, C.Z. and A.M..

DECLARATION OF INTERESTS

A.M., J.B., and M.U. are the founder and shareholders of ScandiBi3o Therapeutics and they filed a patent application on the use of CMA to treat NAFLD patients. The other authors declare no conflict of interest.

Received: July 8, 2021
Revised: September 16, 2021
Accepted: September 30, 2021
Published: November 19, 2021

REFERENCES

Alonso, C., Fernandez-Ramos, D., Varela-Rey, M., Martinez-Arranz, I., Navasa, N., Van Liempd, S.M., Lavin Trueba, J.L., Mayo, R., Ilisso, C.P., de Juan, V.G., et al. (2017). Metabolomic identification of Subtypes of nonalcoholic steatohepatitis. Gastroenterology *152*, 1449–1461.e1447. https://doi.org/10.1053/j.gastro.2017.01.015.

Amid, C., Alako, B.T.F., Balavenkataraman Kadhirvelu, V., Burdett, T., Burgin, J., Fan, J., Harrison, P.W., Holt, S., Hussein, A., Ivanov, E., et al. (2020). The European nucleotide archive in 2019. Nucleic Acids Res. *48*, D70–D76. https://doi.org/10.1093/nar/gkz1063.

Arif, M., Zhang, C., Li, X., Gungor, C., Cakmak, B., Arslanturk, M., Tebani, A., Ozcan, B., Subas, O., Zhou, W., et al. (2021). iNetModels 2.0: an interactive visualization and database of multi-omics data. Nucleic Acids Res. https://doi.org/10.1093/nar/gkab254.

Asrani, S.K., Devarbhavi, H., Eaton, J., and Kamath, P.S. (2019). Burden of liver diseases in the world. J. Hepatol. *70*, 151–171. https://doi.org/10.1016/j.jhep.2018.09.014.

Assenov, Y., Ramirez, F., Schelhorn, S.E., Lengauer, T., and Albrecht, M. (2008). Computing topological parameters of biological networks. Bioinformatics *24*, 282–284. https://doi.org/10.1093/bioinformatics/btm554.

Azzu, V., Vacca, M., Kamzolas, I., Hall, Z., Leslie, J., Carobbio, S., Virtue, S., Davies, S.E., Lukasik, A., Dale, M., et al. (2021). Suppression of insulin-induced gene 1 (INSIG1) function promotes hepatic lipid remodelling and restrains NASH progression. Mol. Metab. *48*, 101210. https://doi.org/10.1016/j.molmet.2021.101210.

Bray, N.L., Pimentel, H., Melsted, P., and Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. Nat. Biotechnol. *34*, 525–527. https://doi.org/10.1038/nbt.3519.

Calabrese, G.M., Mesner, L.D., Stains, J.P., Tommasini, S.M., Horowitz, M.C., Rosen, C.J., and Farber, C.R. (2017). Integrating GWAS and Co-expression network data identifies bone mineral density genes SPTBN1 and MARK3 and an osteoblast functional module. Cell Syst. *4*, 46–59.e44. https://doi.org/10.1016/j.cels.2016.10.014.

Califano, A., Butte, A.J., Friend, S., Ideker, T., and Schadt, E. (2012). Leveraging models of cell regulation and GWAS data in integrative network-based association studies. Nat. Genet. *44*, 841–847. https://doi.org/10.1038/ng.2355.

Cerami, E., Demir, E., Schultz, N., Taylor, B.S., and Sander, C. (2010). Automated network analysis identifies core pathways in glioblastoma. PLoS One *5*, e8918. https://doi.org/10.1371/journal.pone.0008918.

Chella Krishnan, K., Kurt, Z., Barrere-Cain, R., Sabir, S., Das, A., Floyd, R., Vergnes, L., Zhao, Y., Che, N., Charugundla, S., et al. (2018). Integration of multi-omics data from mouse diversity panel highlights mitochondrial dysfunction in non-alcoholic fatty liver disease. Cell Syst. *6*, 103–115.e107. https://doi.org/10.1016/j.cels.2017.12.006.

Chen, J., Bardes, E.E., Aronow, B.J., and Jegga, A.G. (2009). ToppGene Suite for gene list enrichment analysis and candidate gene prioritization. Nucleic Acids Res. *37*, W305–W311. https://doi.org/10.1093/nar/gkp427.

Choobdar, S., Ahsen, M.E., Crawford, J., Tomasoni, M., Fang, T., Lamparter, D., Lin, J., Hescott, B., Hu, X., Mercer, J., et al. (2019). Assessment of network module identification across complex diseases. Nat. Methods *16*, 843–852. https://doi.org/10.1038/s41592-019-0509-5.

Cline, M.S., Smoot, M., Cerami, E., Kuchinsky, A., Landys, N., Workman, C., Christmas, R., Avila-Campilo, I., Creech, M., Gross, B., et al. (2007). Integration of biological networks and gene expression data using cytoscape. Nat. Protoc. *2*, 2366–2382. https://doi.org/10.1038/nprot.2007.324.

Consortium, G.T. (2013). The genotype-tissue expression (GTEx) project. Nat. Genet. *45*, 580–585. https://doi.org/10.1038/ng.2653.

Doncheva, N.T., Assenov, Y., Domingues, F.S., and Albrecht, M. (2012). Topological analysis and interactive visualization of biological networks and protein structures. Nat. Protoc. *7*, 670–685. https://doi.org/10.1038/nprot.2012.004.

Donnelly, K.L., Smith, C.I., Schwarzenberg, S.J., Jessurun, J., Boldt, M.D., and Parks, E.J. (2005). Sources of fatty acids stored in liver and secreted via lipoproteins in patients with nonalcoholic fatty liver disease. J. Clin. Invest. *115*, 1343–1351. https://doi.org/10.1172/JCI23621.

El-Agroudy, N.N., Kurzbach, A., Rodionov, R.N., O'Sullivan, J., Roden, M., Birkenfeld, A.L., and Pesta, D.H. (2019). Are Lifestyle therapies effective for NAFLD treatment? Trends Endocrinol. Metab. *30*, 701–709. https://doi.org/10.1016/j.tem.2019.07.013.

Ericsson, J., Usheva, A., and Edwards, P.A. (1999). YY1 is a negative regulator of transcription of three sterol regulatory element-binding protein-responsive genes. J. Biol. Chem. *274*, 14508–14513. https://doi.org/10.1074/jbc.274.20.14508.

Fagerberg, L., Hallstrom, B.M., Oksvold, P., Kampf, C., Djureinovic, D., Odeberg, J., Habuka, M., Tahmasebpoor, S., Danielsson, A., Edlund, K., et al. (2014). Analysis of the human tissue-specific expression by genome-wide integration of transcriptomics and antibody-based proteomics. Mol. Cell Proteomics *13*, 397–406. https://doi.org/10.1074/mcp.M113.035600.

Friedman, S.L., Neuschwander-Tetri, B.A., Rinella, M., and Sanyal, A.J. (2018). Mechanisms of NAFLD development and therapeutic strategies. Nat. Med. *24*, 908–922. https://doi.org/10.1038/s41591-018-0104-9.

Golabi, P., Fukui, N., Paik, J., Sayiner, M., Mishra, A., and Younossi, Z.M. (2019). Mortality risk detected by atherosclerotic cardiovascular disease score in patients with nonalcoholic fatty liver disease. Hepatol. Commun. *3*, 1050–1060. https://doi.org/10.1002/hep4.1387.

Goldstein, J.L., and Brown, M.S. (2009). The LDL receptor. Arterioscler Thromb. Vasc. Biol. *29*, 431–438. https://doi.org/10.1161/ATVBAHA.108.179564.

Govaere, O., Cockell, S., Tiniakos, D., Queen, R., Younes, R., Vacca, M., Alexander, L., Ravaioli, F., Palmer, J., Petta, S., et al. (2020). Transcriptomic profiling across the nonalcoholic fatty liver disease spectrum reveals gene signatures for steatohepatitis and fibrosis. Sci. Transl Med. *12*. https://doi.org/10.1126/scitranslmed.aba4448.

Han, H., Cho, J.W., Lee, S., Yun, A., Kim, H., Bae, D., Yang, S., Kim, C.Y., Lee, M., Kim, E., et al. (2018). TRRUST v2: an expanded reference database of human and mouse transcriptional regulatory interactions. Nucleic Acids Res. *46*, D380–D386. https://doi.org/10.1093/nar/gkx1013.

Hoang, S.A., Oseini, A., Feaver, R.E., Cole, B.K., Asgharpour, A., Vincent, R., Siddiqui, M., Lawson, M.J., Day, N.C., Taylor, J.M., et al. (2019). Gene expression predicts histological severity and reveals distinct molecular profiles of nonalcoholic fatty liver disease. Sci. Rep. *9*, 12541. https://doi.org/10.1038/s41598-019-48746-5.

Huan, T., Zhang, B., Wang, Z., Joehanes, R., Zhu, J., Johnson, A.D., Ying, S., Munson, P.J., Raghavachari, N., Wang, R., et al. (2013). A systems biology framework identifies molecular underpinnings of coronary heart disease. Arterioscler Thromb. Vasc. Biol. *33*, 1427–1434. https://doi.org/10.1161/ATVBAHA.112.300112.

Huang, D.Q., El-Serag, H.B., and Loomba, R. (2021). Global epidemiology of NAFLD-related HCC: trends, predictions, risk factors and prevention. Nat. Rev. Gastroenterol. Hepatol. *18*, 223–238. https://doi.org/10.1038/s41575-020-00381-6.

Huang, M., Bao, J., Hallstrom, B.M., Petranovic, D., and Nielsen, J. (2017). Efficient protein production by yeast requires global tuning of metabolism. Nat. Commun. *8*, 1131. https://doi.org/10.1038/s41467-017-00999-2.

Ioannou, G.N. (2016). The role of cholesterol in the pathogenesis of NASH. Trends Endocrinol. Metab. *27*, 84–95. https://doi.org/10.1016/j.tem.2015.11.008.

Ioannou, G.N., Green, P., Kerr, K.F., and Berry, K. (2019). Models estimating risk of hepatocellular carcinoma in patients with alcohol or NAFLD-related cirrhosis for risk stratification. J. Hepatol. *71*, 523–533. https://doi.org/10.1016/j.jhep.2019.05.008.

Jensen, T., Abdelmalek, M.F., Sullivan, S., Nadeau, K.J., Green, M., Roncal, C., Nakagawa, T., Kuwabara, M., Sato, Y., Kang, D.H., et al. (2018). Fructose and sugar: a major mediator of non-alcoholic fatty liver disease. J. Hepatol. *68*, 1063–1075. https://doi.org/10.1016/j.jhep.2018.01.019.

Jiao, N., Baker, S.S., Chapa-Rodriguez, A., Liu, W., Nugent, C.A., Tsompana, M., Mastrandrea, L., Buck, M.J., Baker, R.D., Genco, R.J., et al. (2018). Suppressed hepatic bile acid signalling despite elevated production of primary and secondary bile acids in NAFLD. Gut *67*, 1881–1891. https://doi.org/10.1136/gutjnl-2017-314307.

Kuleshov, M.V., Jones, M.R., Rouillard, A.D., Fernandez, N.F., Duan, Q., Wang, Z., Koplev, S., Jenkins, S.L., Jagodnik, K.M., Lachmann, A., et al. (2016). Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. Nucleic Acids Res. *44*, W90–W97. https://doi.org/10.1093/nar/gkw377.

Lee, S., Zhang, C., Arif, M., Liu, Z., Benfeitas, R., Bidkhori, G., Deshmukh, S., Al Shobky, M., Lovric, A., Boren, J., et al. (2018). TCSBN: a database of tissue and cancer specific biological networks. Nucleic Acids Res. *46*, D595–D600. https://doi.org/10.1093/nar/gkx994.

Lee, S., Zhang, C., Kilicarslan, M., Piening, B.D., Bjornson, E., Hallstrom, B.M., Groen, A.K., Ferrannini, E., Laakso, M., Snyder, M., et al. (2016). Integrated network analysis reveals an association between plasma mannose levels and insulin resistance. Cell Metab. *24*, 172–184. https://doi.org/10.1016/j.cmet.2016.05.026.

Lewis, J.E., and Kemp, M.L. (2021). Integration of machine learning and genome-scale metabolic modeling identifies multi-omics biomarkers for radiation resistance. Nat. Commun. *12*, 2700. https://doi.org/10.1038/s41467-021-22989-1.

Liu, D., Wong, C.C., Fu, L., Chen, H., Zhao, L., Li, C., Zhou, Y., Zhang, Y., Xu, W., Yang, Y., et al. (2018a). Squalene epoxidase drives NAFLD-induced hepatocellular carcinoma and is a pharmaceutical target. Sci. Transl Med. *10*. https://doi.org/10.1126/scitranslmed.aap9840.

Liu, J., Lichtenberg, T., Hoadley, K.A., Poisson, L.M., Lazar, A.J., Cherniack, A.D., Kovatich, A.J., Benz, C.C., Levine, D.A., Lee, A.V., et al. (2018b). An integrated TCGA pan-cancer clinical data resource to drive high-quality survival outcome analytics. Cell *173*, 400–416.e411. https://doi.org/10.1016/j.cell.2018.02.052.

Liu, Z., Zhang, C., Lee, S., Kim, W., Klevstig, M., Harzandi, A.M., Sikanic, N., Arif, M., Stahlman, M., Nielsen, J., et al. (2019). Pyruvate kinase L/R is a regulator of lipid metabolism and mitochondrial function. Metab. Eng. *52*, 263–272. https://doi.org/10.1016/j.ymben.2019.01.001.

Loomba, R., Friedman, S.L., and Shulman, G.I. (2021). Mechanisms and disease consequences of nonalcoholic fatty liver disease. Cell *184*, 2537–2564. https://doi.org/10.1016/j.cell.2021.04.015.

Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. *15*, 550. https://doi.org/10.1186/s13059-014-0550-8.

Lu, Y., Ma, Z., Zhang, Z., Xiong, X., Wang, X., Zhang, H., Shi, G., Xia, X., Ning, G., and Li, X. (2014). Yin Yang 1 promotes hepatic steatosis through repression of farnesoid X receptor in obese mice. Gut *63*, 170–178. https://doi.org/10.1136/gutjnl-2012-303150.

Mardinoglu, A., Agren, R., Kampf, C., Asplund, A., Uhlen, M., and Nielsen, J. (2014). Genome-scale metabolic modelling of hepatocytes reveals serine deficiency in patients with non-alcoholic fatty liver disease. Nat. Commun. *5*, 3083. https://doi.org/10.1038/ncomms4083.

Mardinoglu, A., Boren, J., Smith, U., Uhlen, M., and Nielsen, J. (2018). Systems biology in hepatology: approaches and applications. Nat. Rev. Gastroenterol. Hepatol. *15*, 365–377. https://doi.org/10.1038/s41575-018-0007-8.

Min, H.K., Kapoor, A., Fuchs, M., Mirshahi, F., Zhou, H., Maher, J., Kellum, J., Warnick, R., Contos, M.J., and Sanyal, A.J. (2012). Increased hepatic synthesis and dysregulation of cholesterol metabolism is associated with the severity of nonalcoholic fatty liver disease. Cell Metab. *15*, 665–674. https://doi.org/10.1016/j.cmet.2012.04.004.

Miryala, S.K., and Ramaiah, S. (2019). Exploring the multi-drug resistance in Escherichia coli O157:H7 by gene interaction network: a systems biology approach. Genomics *111*, 958–965. https://doi.org/10.1016/j.ygeno.2018.06.002.

Mullard, A. (2020). FDA rejects NASH drug. Nat. Rev. Drug Discov. *19*, 501. https://doi.org/10.1038/d41573-020-00126-9.

Musso, G., Gambino, R., and Cassader, M. (2013). Cholesterol metabolism and the pathogenesis of non-alcoholic steatohepatitis. Prog. Lipid Res. *52*, 175–191. https://doi.org/10.1016/j.plipres.2012.11.002.

Natesampillai, S., Fernandez-Zapico, M.E., Urrutia, R., and Veldhuis, J.D. (2006). A novel functional interaction between the Sp1-like protein KLF13 and SREBP-Sp1 activation complex underlies regulation of low density lipoprotein receptor promoter function. J. Biol. Chem. *281*, 3040–3047. https://doi.org/10.1074/jbc.M509417200.

Nayak, R.R., Kearns, M., Spielman, R.S., and Cheung, V.G. (2009). Coexpression network based on natural variation in human gene expression reveals gene interactions and functions. Genome Res. *19*, 1953–1962. https://doi.org/10.1101/gr.097600.109.

Newsome, P.N., Buchholtz, K., Cusi, K., Linder, M., Okanoue, T., Ratziu, V., Sanyal, A.J., Sejling, A.S., Harrison, S.A., and Investigators, N.N. (2021). A placebo-controlled trial of subcutaneous semaglutide in nonalcoholic steatohepatitis. N. Engl. J. Med. *384*, 1113–1124. https://doi.org/10.1056/NEJMoa2028395.

Oliveira, A.P., Patil, K.R., and Nielsen, J. (2008). Architecture of transcriptional regulatory circuits is knitted over the topology of bio-molecular interaction networks. BMC Syst. Biol. *2*, 17. https://doi.org/10.1186/1752-0509-2-17.

Parkinson, H., Sarkans, U., Shojatalab, M., Abeygunawardena, N., Contrino, S., Coulson, R., Farne, A., Lara, G.G., Holloway, E., Kapushesky, M., et al. (2005). ArrayExpress—a public repository for microarray gene expression data at the EBI. Nucleic Acids Res. *33*, D553–D555. https://doi.org/10.1093/nar/gki056.

Patil, K.R., and Nielsen, J. (2005). Uncovering transcriptional regulation of metabolism by using metabolic network topology. Proc. Natl. Acad. Sci. U. S. A. *102*, 2685–2689. https://doi.org/10.1073/pnas.0406811102.

Perdomo, C.M., Fruhbeck, G., and Escalada, J. (2019). Impact of nutritional changes on nonalcoholic fatty liver disease. Nutrients *11*, 677. https://doi.org/10.3390/nu11030677.

Pinero, J., Bravo, A., Queralt-Rosinach, N., Gutierrez-Sacristan, A., Deu-Pons, J., Centeno, E., Garcia-Garcia, J., Sanz, F., and Furlong, L.I. (2017). DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. Nucleic Acids Res. *45*, D833–D839. https://doi.org/10.1093/nar/gkw943.

Pinero, J., Ramirez-Anguita, J.M., Sauch-Pitarch, J., Ronzano, F., Centeno, E., Sanz, F., and Furlong, L.I. (2020). The DisGeNET knowledge platform for disease genomics: 2019 update. Nucleic Acids Res. *48*, D845–D855. https://doi.org/10.1093/nar/gkz1021.

Ray, K. (2018). NAFLD-HCC: target cholesterol. Nat. Rev. Gastroenterol. Hepatol. *15*, 390. https://doi.org/10.1038/s41575-018-0029-2.

Saeed, E. (2021). Core liver homeostatic co-expression networks are preserved but respond to perturbations in an organism- and disease-specific manne. Cell Syst. https://doi.org/10.1016/j.cels.2021.04.004.

Shen, T., Lu, Y., and Zhang, Q. (2020). High squalene epoxidase in tumors predicts worse survival in patients with hepatocellular carcinoma: integrated bioinformatic analysis on NAFLD and HCC. Cancer Control *27*. 1073274820914663. https://doi.org/10.1177/1073274820914663.

Stower, H. (2021). Therapy for NASH. Nat. Med. *27*, 21. https://doi.org/10.1038/s41591-020-01219-z.

Traag, V.A., Waltman, L., and van Eck, N.J. (2019). From Louvain to Leiden: guaranteeing well-

connected communities. Sci. Rep. *9*, 5233. https://doi.org/10.1038/s41598-019-41695-z.

Uhlen, M., Fagerberg, L., Hallstrom, B.M., Lindskog, C., Oksvold, P., Mardinoglu, A., Sivertsson, A., Kampf, C., Sjostedt, E., Asplund, A., et al. (2015). Proteomics. Tissue-based map of the human proteome. Science *347*, 1260419. https://doi.org/10.1126/science.1260419.

Uhlen, M., Zhang, C., Lee, S., Sjostedt, E., Fagerberg, L., Bidkhori, G., Benfeitas, R., Arif, M., Liu, Z., Edfors, F., et al. (2017). A pathology atlas of the human cancer transcriptome. Science *357*. https://doi.org/10.1126/science.aan2507.

van Dam, S., Vosa, U., van der Graaf, A., Franke, L., and de Magalhaes, J.P. (2018). Gene co-expression analysis for functional classification and gene-disease predictions. Brief Bioinform *19*, 575–592. https://doi.org/10.1093/bib/bbw139.

Varemo, L., Nielsen, J., and Nookaew, I. (2013). Enriching the gene set analysis of genome-wide data by incorporating directionality of gene expression and combining statistical hypotheses and methods. Nucleic Acids Res. *41*, 4378–4391. https://doi.org/10.1093/nar/gkt111.

Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., et al. (2020). SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat. Methods *17*, 261–272. https://doi.org/10.1038/s41592-019-0686-2.

Vivian, J., Rao, A.A., Nothaft, F.A., Ketchum, C., Armstrong, J., Novak, A., Pfeil, J., Narkizian, J., Deran, A.D., Musselman-Brown, A., et al. (2017). Toil enables reproducible, open source, big biomedical data analyses. Nat. Biotechnol. *35*, 314–316. https://doi.org/10.1038/nbt.3772.

Wainberg, M., Kamber, R.A., Balsubramani, A., Meyers, R.M., Sinnott-Armstrong, N., Hornburg, D., Jiang, L., Chan, J., Jian, R., Gu, M., et al. (2021). A genome-wide atlas of co-essential modules assigns function to uncharacterized genes. Nat. Genet. *53*, 638–649. https://doi.org/10.1038/s41588-021-00840-z.

Wu, G.Y., Rui, C., Chen, J.Q., Sho, E., Zhan, S.S., Yuan, X.W., and Ding, Y.T. (2017). MicroRNA-122 inhibits lipid droplet formation and hepatic triglyceride accumulation via Yin Yang 1. Cell Physiol. Biochem. *44*, 1651–1664. https://doi.org/10.1159/000485765.

Ye, Q., Zou, B., Yeo, Y.H., Li, J., Huang, D.Q., Wu, Y., Yang, H., Liu, C., Kam, L.Y., Tan, X.X.E., et al. (2020). Global prevalence, incidence, and outcomes of non-obese or lean non-alcoholic fatty liver disease: a systematic review and meta-analysis. Lancet Gastroenterol. Hepatol. *5*, 739–752. https://doi.org/10.1016/S2468-1253(20)30077-7.

Younossi, Z., Anstee, Q.M., Marietti, M., Hardy, T., Henry, L., Eslam, M., George, J., and Bugianesi, E. (2018). Global burden of NAFLD and NASH: trends, predictions, risk factors and prevention. Nat. Rev. Gastroenterol. Hepatol. *15*, 11–20. https://doi.org/10.1038/nrgastro.2017.109.

Younossi, Z.M., Golabi, P., de Avila, L., Paik, J.M., Srishord, M., Fukui, N., Qiu, Y., Burns, L., Afendy, A., and Nader, F. (2019). The global epidemiology of NAFLD and NASH in patients with type 2 diabetes: a systematic review and meta-analysis. J. Hepatol. *71*, 793–801. https://doi.org/10.1016/j.jhep.2019.06.021.

Younossi, Z.M., Koenig, A.B., Abdelatif, D., Fazel, Y., Henry, L., and Wymer, M. (2016). Global epidemiology of nonalcoholic fatty liver disease-meta-analytic assessment of prevalence, incidence, and outcomes. Hepatology *64*, 73–84. https://doi.org/10.1002/hep.28431.

Yu, N.Y., Hallstrom, B.M., Fagerberg, L., Ponten, F., Kawaji, H., Carninci, P., Forrest, A.R., Fantom, C., Hayashizaki, Y., Uhlen, M., and Daub, C.O. (2015). Complementing tissue characterization by integrating transcriptome profiling from the Human Protein Atlas and from the FANTOM5 consortium. Nucleic Acids Res. *43*, 6787–6798. https://doi.org/10.1093/nar/gkv608.

Zhang, C., Bjornson, E., Arif, M., Tebani, A., Lovric, A., Benfeitas, R., Ozcan, M., Juszczak, K., Kim, W., Kim, J.T., et al. (2020). The acute effect of metabolic cofactor supplementation: a potential therapeutic strategy against non-alcoholic fatty liver disease. Mol. Syst. Biol. *16*, e9495. https://doi.org/10.15252/msb.209495.

Zhang, X., Coker, O.O., Chu, E.S., Fu, K., Lau, H.C.H., Wang, Y.X., Chan, A.W.H., Wei, H., Yang, X., Sung, J.J.Y., and Yu, J. (2021). Dietary cholesterol drives fatty liver-associated liver cancer by modulating gut microbiota and metabolites. Gut *70*, 761–774. https://doi.org/10.1136/gutjnl-2019-319664.

STAR★ METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
NAFLD cohort 1 samples	The European Nucleotide Archive	SRP217231 https://www.ebi.ac.uk/ena/
NAFLD cohort 2 samples	The European Nucleotide Archive	SRP197353 https://www.ebi.ac.uk/ena/
NAFLD cohort 3 samples	ArrayExpress Archive	E-MTAB-9815 https://www.ebi.ac.uk/arrayexpress/
Mouse NAFLD model	This paper	GSE184019 https://www.ncbi.nlm.nih.gov/geo/
Experimental models: Organisms/strains		
C57BL/6J mice	Gothenburg, SE	NA
Software and algorithms		
R language version 4.0.3		https://cran.r-project.org/
Matlab language version R2020b		https://www.mathworks.com/
Cytoscape version 3.8.2		https://cytoscape.org
Python version 3.8	Python Software Foundation	https://www.python.org
Kallisto	Bray et al. (2016)	https://pachterlab.github.io/kallisto/

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contract, Adil Mardinoglu (adilm@scilifelab.se).

Materials availability

This study did not generate new unique reagents.

Data and code availability

The raw expression data derived from mouse samples have been reposited at Gene Expression Omnibus and are publicly available as of the date of publication. Accession numbers are listed in the Key resources table.

The paper does not report original code.

Any additional information required to reanalyze the data reported in this paper is available from the lead contract upon request.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Nine C57BL/6J mice were fed a standard mouse chow diet and housed in a 12-h light–dark cycle. From the age of 8 weeks, the mice were then divided into two groups of 5 mice fed with chow diet, 4 mice fed with high-sucrose diet for 3 weeks, respectively.

METHOD DETAILS

Transcriptomics data from mouse model

Nine C57BL/6J mice were fed a standard mouse chow diet and housed in a 12-h light–dark cycle. From the age of 8 weeks, the mice were then divided into two groups of 5 mice fed with chow diet, 4 mice fed with high-sucrose diet for 3 weeks, respectively. At the age of 11 weeks, all mice are sacrificed and liver necropsy were taken for RNA sequencing. RNA sequencing library were prepared with Illumina RNA-Seq with Poly-A selections. Subsequently, the libraries were sequenced on NovaSeq6000 (NovaSeq Control Software 1.6.0/

RNA v3.4.4) with a 2×51 setup using 'NovaSeqXp' workflow in 'S1' mode flow cell. The Bcl was converted to fastq by bcl2fastq_v2.19.1.403 from CASAVA software suite (Sanger/phred33/Illumina 1.8+ quality scale). The fastq files for mice were then processed using a standard protocol of Kallisto (Bray et al., 2016).

QUANTIFICATION AND STATISTICAL ANALYSIS

Data retrieving and pre-processing

Each dataset was pre-processed independently:

NAFLD cohorts. hepatic RNA-seq (raw fastq files) of NAFLD cohort 1 (Govaere et al., 2020) and cohort 2 (Hoang et al., 2019) were retrieved from European Nucleotide Archive (ENA) database (https://www.ebi.ac.uk/ena/) (Amid et al., 2020) under accession numbers SRP217231 (215 biopsy-proven NAFLD patients) and SRP197353 (78 biopsy-proven NAFLD patients), respectively; Hepatic RNA-seq of NAFLD cohort 3 (Azzu et al., 2021) with 58 biopsy-proven NAFLD patients were retrieved from the ArrayExpress data repository (Parkinson et al., 2005) under accession number E-MTAB-9815. Principle component analysis (PCA) was used to detect outlier samples (Figure S1) and three outlier samples in NAFLD cohort 2 were removed based on this analysis. Afterwards, gene abundance in both transcripts per million (TPMs)) and raw count were quantified using the Kallisto (Bray et al., 2016) pipeline based on human genome (ensemble 102 version). We subsequently used DESeq2 R package following a standard protocol (Love et al., 2014) to identify differentially expressed genes (DEGs, adjusted p-value < 0.01) and performed KEGG pathway enrichment using the Platform for Integrative Analysis of Omics (PIANO) R package (Varemo et al., 2013).

GTEx cohort. The RNA-seq data with gene abundance in transcript TPMs from human tissues was retrieved from Genotype tissue expression (GTEx) (https://gtexportal.org/home/datasets) (Consortium, 2013) and retained the samples with available dataset in liver tissue.

Construction of co-expression network and analysis

Considering the dramatic increase in size owing to the many gene isoforms and non-coding RNAs (van Dam et al., 2018), we used the "protein-coding genes" for annotation of RNA-seq dataset and then constructed the co-expression network in gene level. For each dataset, we first filtered out lowly-expressed genes based on their median gene expression level (TPM <1) and constructed co-expression networks by generating gene pairs Spearman correlation ranks within liver tissue, which was performed using "spearmanr" function from *SciPy* (Virtanen et al., 2020) in Python 3.8. Next, considering the network with negative correlation has relatively low correlation scores, we retained top 10% positively correlated genes that fulfilled FDR < 0.05 on the network (Arif et al., 2021) and performed module detection analysis using Leiden algorithm (Traag et al., 2019), implemented by Python package *leidenalg* (version 0.7.0) with "ModularityVertexPartition" to find the optimal partition. Modules with less than 30 genes were discarded to be able to get significant functional analysis results in the downstream analysis.

To explore the module similarity between different cohorts, we calculated the Jaccard index, which is simply defined as the size of the intersection between two modules divided by the size of the union of the same two modules, and used hypergeometric test to determine whether the genes in one module significantly overlapped with the genes in another module. The overlap was considered as significant when p-value less than 0.05. Topological and node properties of modules were determined using the *NetworkAnalyzer* (Assenov et al., 2008) plugin implemented in *Cytoscape* (version, 3.8.2) (Cline et al., 2007).

Functional annotation of modules in co-expression network of cohort

KEGG enrichment analysis. We performed functional enrichment analysis for the gene lists of each module of co-expression network using hypergeometric test, which is implemented by the python package *gseapy* (version 0.9.16; https://github.com/zqfang/gseapy), all gene sets of KEGG pathway were obtained from database source of *Enrichr* (Kuleshov et al., 2016).

Disease enrichment analysis. DisGeNet (Pinero et al., 2017) is a platform integrating information of gene-disease association from several public data sources and literature. In our analysis, the lists of diseases enriched by the gene lists in each network module were retrieved from the DisGeNet database using ToppFun of the ToppGene suite (Chen et al., 2009), all gene sets in detected modules were used as

background gene sets. Disease terms with Benjamini-Hochberg corrected p-value <0.05 were retained and top 20 for each disease-associated module were presented.

Transcription factor enrichment analysis. We retrieved the human Transcriptional Regulatory Relationship Unravelled by Sentence-based Text mining (TRRUST) v2 database (https://www.grnpedia.org/trrust/) (Han et al., 2018) and obtained the lists of transcription factor and associated targets, which derived from 7,148 PubMed articles in where small-scale experimental studies of transcriptional regulation were described. In total, 9,395 TF-target regulatory relationships of 795 TFs and 2,493 targets were supplied as database for *Enrichr* (Kuleshov et al., 2016), implemented by the python package *gseapy* (version 0.9.16; https://github.com/zqfang/gseapy).

Reporter metabolite and reporter transcription factor analyses

To investigate the detailed metabolic differences associated with NAFLD, we first performed reporter metabolites analysis (Varemo et al., 2013) using the PIANO R package with topological information from liver-specific GEM *iHepatocytes2322* (Mardinoglu et al., 2014). Differential expression level of genes (log2-fold change) in each contrast and corresponding significant levels (p value) were used as input.

To validate the enriched transcription factors in disease-associated modules, we also employed the PIANO R package to perform reporter transcription factor (TF) analysis (Huang et al., 2017; Varemo et al., 2013) in which log2-fold change and p-value of genes, as well as transcriptional regulatory information of TF-target from TRRUST database (Han et al., 2018) were used as input. In the reporter TF analysis of a module, we kept the original p-value of genes in the module and assigned the p-value of genes that are not in the module to 1 in order to eliminate influences from genes in other modules.

TCGA data process and survival analysis

The transcript-expression level profiles (TPM) had been downloaded from Toil (Vivian et al., 2017) under the project ID of TCGA-LIHC. We screened all samples in TCGA-LIHC cohorts and kept 363 donors with both primary tumour solid tissue samples and clinical information. We only extracted tumour samples with identifier "A" for liver hepatocellular carcinoma and subsequently quantified the mRNA expression by Kallisto (Bray et al., 2016) based on the GENCODE reference transcriptome (version 23). Genes with an average TPM>1 were reserved for the following analysis. The clinical information was collected from TCGA Pan-Cancer Clinical Data Resource (TCGA-CDR) (Liu et al., 2018b). Samples with a survival time of zero-day were excluded.

To investigate if expression level of a gene is associated with patient outcomes, we first divided samples into high and low expression groups based on TPM value of the gene. Next, we performed Kaplan-Meier survival analysis to determine the association; the survival outcomes were then compared based on log-rank tests. To choose the best TPM cut-offs for grouping, all TPM values from the 20th to 80th percentiles were used to group the patients. Significant differences in the survival outcomes of the groups were examined, and the value with the lowest log-rank P-value is selected. The R package "survival" and graphics "ggplot" was used during the Kaplan-Meier analysis. Genes with log-rank P values less than 0.05 were defined as prognostic genes. In addition, if the group of patients with high expression of a selected prognostic gene has a higher observed event than the expected event, it is an unfavourable prognostic gene; otherwise, it is a favourable prognostic gene. All analysis were conducted with R.